The DDBJ/EMBL/GenBank® Feature Table: Definition

Version 1.03
August 6, 1990

This version incorporates minor corrections to the GenBank new format feature table examples in Section 5 and Appendix I. In addition, this version omits certain reference appendices included in previous versions.

DNA Data Bank of Japan, Mishima, Japan
EMBL Data Library, Heidelberg, Federal Republic of Germany
GenBank, Los Alamos, NM and Mountain View, CA, USA

Contents

The DDBJ/EMBL/GenBank®
Feature Table:Definition

Version 1.03
August 6, 1990

1 Introduction

Nucleic acid sequences provide the fundamental starting point for describing and understanding the structure, function, and development of genetically diverse organisms. The primary structure of RNAs and proteins can be derived directly from the DNA or RNA genomic sequences. However, the translation, expression dependencies, and temporal sequence of events are neither straightforward nor well-defined. The GenBank, EMBL, and DDBJ nucleic acid sequence data banks have from their inception used tables of sites and features to describe the roles and locations of higher order sequence domains and elements within the genome of an organism. The independent development of feature table formats and annotation standards at EMBL and GenBank (later adopted by DDBJ) created significant difficulties for the data banks' data sharing efforts. As a result, in February, 1986, GenBank and EMBL began a collaborative effort (joined by DDBJ in 1987) to devise a common feature table format and common standards for annotation practice.

Early in this collaborative process it was recognized that both existing representational schemes suffered from similar inadequacies:

- Much of the information contained in the tables was difficult or impossible to extract by automatic methods.
- Sequences or features now recognized as important could not be annotated, but the tables' syntaxes limited their extension to include new classes of features.
- Features were not citable. They had no unique identifiers; hence, there could be no mechanism for linking between databases.
- The tables' syntaxes severely limited mechanisms for parsibly expressing complex features such as alternate splicing, circular molecules, read-through stop codons, sequence variation, multiple reading frames, secondary nucleic acid structures, pseudogenes, and other complex relationships between sequence regions.

2 Overview of the new Feature Table format

The overall goal of the new feature table design is to provide a more extensive vocabulary for describing features in a flexible framework for manipulating them. The range of features to be represented is diverse, including regions which:

(a) perform a biological function,
(b) affect or are the result of the expression of a biological function,
(c) interact with other molecules,
(d) affect replication of a sequence,
(e) affect or are the result of recombination of different sequences,
(f) are a recognizable repeated unit,
(g) have secondary or tertiary structure,
(h) exhibit variation, or
(i) have been revised or corrected.

The new feature table expands the feature vocabulary and adds new data items to allow more flexibility and a clearer specification of complex features. The format design, which is based on a tabular approach consists of the following items:

1. Feature key	a keyword indicating functional group
2. Location	instructions for finding the feature
3. Qualifiers	auxiliary information about a feature

Each item will be discussed in more detail later; however, this design alleviates the limitations mentioned in the Introduction in the following ways:

- Features are distinct, citable entities
  A new feature table data item -- the feature label (specified with a qualifier) -- permits direct identification of a feature. Therefore a feature can be referred to within an entry by feature label, between entries by primary accession number and feature label and between databases by database name, primary accession number and feature label. Note that this labelling convention allows cross-referencing to other databases (e.g., protein and genetic mapping databases as well as other nucleotide sequence databases) which adopt a compatible scheme of accession numbers and feature labels.

- A much richer set of keys allows specific annotation of important sequence features.
  Numerous features of biological significance (e.g., TATA and CAAT boxes; untranslated regions) which previously did not have distinct feature keys can now be clearly indicated.

- Related features can be easily specified and retrieved.
  Feature keys are arranged hierarchically, allowing complex and compound features to be expressed. Both location operators and the feature keys show feature relationships even when the features are not contiguous. Separate features, such as multiple mutations in a single strain, can be tied together using the "group" location operator. The hierarchy of feature keys allows broad categories of biological functionality, such as rRNAs, to be easily retrieved.

- Generic feature keys provide a means for entering new or undefined features.
  A number of "generic" or miscellaneous feature keys have been added to permit annotation of features that cannot be adequately described by existing feature keys. These generic feature keys will serve as an intermediate step in the identification and addition of new feature keys. The syntax has been designed to allow the addition of new feature keys as they are required.

- More complex locations (fuzzy and alternate ends, for example) can now be specified.
  Each end point of a feature may be specified as a single point, an alternate set of possible end points, a base number beyond which the end point lies, or a region which contains the end point.

- Features can be combined and manipulated in many different ways.
  The new location field can contain operators or functional descriptors specifying what must be done to the sequence to reproduce the feature. For example, a series of exons may be "join"ed into a full coding sequence.

- Precision and parsibility of descriptive details is greatly enhanced by the provision of standardized qualifiers.
  Information which would have previously been expressible only in the free-text description field can in many cases now be provided as a combination of standardized qualifiers and their controlled-vocabulary values.

- The nature of supporting evidence for a feature now can be explicitly indicated.
  Features, such as open reading frames or sequences showing sequence similarity to consensus sequences, for which there is no direct experimental evidence can be annotated. Therefore, the feature table can incorporate contributions from researchers doing computational analysis of the sequence databases. However, all features that are supported by experimental data will be clearly marked as such.

- The table syntax has been designed to be machine parsible.
  A consistent syntax will allow machine extraction and manipulation of sequences coding for all features in the table.

In addition to addressing known limitations in the current feature table, there were a number of other design goals. The flat file distribution format of the feature table must be easily read and understood; therefore, the format and wording in the new feature table use common biological research terminology whenever possible. For example, an item in the new feature table such as:

```
Key             Location/Qualifiers
CDS             23..400
                /product="alcohol dehydrogenase" /gene="adhI"
                /label=adh1
```

might be read as:

> *The feature called 'adh1', which is a coding sequence beginning at base 23 and ending at base 400, has a product called 'alcohol dehydrogenase' and corresponds to the gene called 'adhI'.*

This relatively straightforward way of reading an item works even for more complex descriptions such as:

```
Key             Location/Qualifiers
protein_bind    one-of((10..21),(15..26))
                /bound_moiety="repressor"
```

which might be read as:

> *This feature (unlabelled) is a protein binding site which binds one of two regions -- either bases 10 to 21 or bases 15 to 26 -- to a repressor.*

(If the repressor sequence were contained in the database, the primary accession number and feature label or base range would be specified instead of 'repressor'.)

The following sections contain detailed explanations of the new feature table design showing conventions for each component of the new feature table, examples of how the format might be implemented, a description of the exact column placement of all the data items and examples of complete sequence entries that have been annotated using the new format. The last section of this document describes known limitations of the current feature table design.

Appendix I gives an example database entry for both the EMBL and GenBank/DDBJ formats. Appendix II describes the format in Backus-Naur-Form (BNF). Appendices III and IV provide reference manuals for the feature table keys and qualifiers, respectively.

This document defines the syntax and vocabulary of the feature table. The syntax is sufficiently flexible to allow expression of a single biological entity in numerous ways. In such cases, the

annotation staffs at the databases will propose conventions for standard means of denoting the entities. These conventions will be contained in The Feature Table Annotation Standards Guide, which will also contain detailed descriptions and examples of the format for each feature key.

This feature table format will be shared by GenBank, EMBL and DDBJ. Comments, corrections, and suggestions may be submitted to any of the database staffs. New format specifications will be added as needed.

3 Feature table components and format

3.1 Naming conventions

Feature table components, including feature keys, qualifiers, accession numbers, database name abbreviations, feature labels, and location operators, are all named following the same conventions. Component names may be no more than 15 characters long and must contain at least one letter. While case should not be regarded as significant in comparing feature labels ('Prot1' and 'pROT1' are the same), the databanks will preserve the case of labels as originally annotated. The following characters are permitted to occur in feature table component names:

Upper-case letters (A-Z)
Lower-case letters (a-z)
Numbers (0-9)
Underscore (_)
Hyphen (-)
Single quotation mark or apostrophe (')
Asterisk (*)

3.2 Feature keys

3.2.1 Purpose

Feature keys indicate (1) the biological nature of the annotated feature or (2) information about changes to or other versions of the sequence. The feature key permits a user to quickly find or retrieve similar features or features with related functions.

3.2.2 Format and conventions

There is a defined list of allowable feature keys which is shown in Appendix III. Each feature must contain a feature key. Features created solely as location references should use a single hyphen (-) as their feature key.

Users may define their own feature keys for local use but user-defined keys' names should begin with an asterisk (*) and follow the naming conventions described in section 3.1. The data banks will, in the future, create new feature keys as necessary. However, key names beginning with an asterisk will be reserved for local use.

3.2.3 Key groups and hierarchy

The feature keys fall into families which are in some sense similar in function and which are annotated in a similar manner. A functional family may have a "generic" or miscellaneous key, which can be recognized by the 'misc.' prefix, that can used for instances not covered by the other defined keys of that group.

The feature key groups are listed below with a short definition and an annotation example:

1) Difference and change features. Indicate ways in which a sequence should be changed to produce a different "version":

```
misc_difference    replace(location,location)
```

2) Expression signal features. Indicate regions containing a signal that alters a biological function:

```
misc_signal        location
```

3) Transcript features. Indicate products made by a region:

```
misc_RNA           location
```

4) Binding features. Indicate that a sequence or nucleotide is covalently, non-covalently or otherwise bound to something else:

```
misc_binding       location
/bound_moiety="bound molecule"
```

5) Repeat features. Indicate repetitive sequence elements:

```
repeat_region      location
```

6) Recombination features. Indicate regions that have been either inserted or deleted by recombination:

```
misc_recomb        location
```

7) Structure features. Indicate sequence for which there is secondary or tertiary structural information:

```
misc_structure     location
```

In addition to the functional groupings shown above, the feature keys can also be arranged in a hierarchical tree based on the degree of specificity or level of detail known about a feature. This hierarchy is shown in outline form in Appendix III where the most general level is the 'misc_feature' key and other keys are arranged in increasing level of detail. By using more general keys, features can be annotated even if their biological functions are insufficiently well characterized to assign them more specific keys.

3.2.4 Feature key examples

Key	Description
conflict	Separate determinations of the "same" sequence differ
rep_origin	Origin of replication
protein_bind	Protein binding site on DNA
CDS	Protein-coding sequence
misc_RNA	Generic label for an undefined RNA
insertion_seq	Insertion element
D-loop	Mitochondrial or other D-loop structure

See Appendix III for descriptions of all feature keys.

3.3 Qualifiers

3.3.1 Purpose

Qualifiers provide a general mechanism for supplying information about features in addition to that conveyed by the key and location.

3.3.2 Format and conventions

Qualifiers take the form of a slash (/) followed by the qualifier name and, if applicable, an equal sign (=) and a value. Values may be: (a) a simple single value, (b) multiple values (of the same data type) separated by commas and enclosed in parentheses, or (c) a list of "tagged" values (value identifiers followed by a colon (:) and the values; the list must be enclosed in parentheses).

If the location descriptor does not need a continuation line, the first qualifier begins a new line in the feature location column. If the location descriptor requires a continuation line, the first qualifier may follow immediately after the location. Any necessary continuation lines begin in the same column. See Section 4 for a complete description of data item positions.

3.3.3 Qualifier values

Since qualifiers convey many different types of information, there are several value formats:

(1) Free text
(2) Controlled vocabulary or enumerated values
(3) Citation or reference numbers
(4) Sequences
(5) Feature labels

3.3.3.1 Free text

Most qualifier values will be a descriptive text phrase which must be enclosed in double quotation marks. When the text occupies more than one line, a single set of quotation marks are required at the beginning and at the end of the text. The text itself may be composed of any printable characters (ASCII values 32-126 decimal). If double quotation marks are used within a free text string, each set (") must be 'escaped' by placing a second double quotation mark immediately before it (""). For example:

/note="This is an example of ""escaped"" quotation marks"

3.3.3.2 Controlled vocabulary or enumerated values

Some qualifiers require values from a controlled vocabulary and can be entered without quotation marks. For example, the '/direction' qualifier has only three values: 'left', 'right' or 'both'. Qualifier value controlled vocabularies, like feature table component names, must be treated as completely case insensitive: they may be entered and displayed in any combination of upper and lower case ('/direction=Left' '/direction=left' and '/direction=LEFT' are all legal and all convey the same meaning). The database staffs reserve the right to regularize the case of qualifier values in the interest of readability, unlike the case of feature labels where the databases will maintain the case as originally entered (see Section 3.4.2). Qualifier value

controlled vocabularies will be maintained by the cooperating database staffs. Examples of controlled vocabularies can be found in Appendices IV and VII. The database staff should be contacted for the current lists.

3.3.3.3 Citation or reference numbers

The citation or published reference number (as enumerated in the entry 'REFERENCE' or 'RN' data item) should be enclosed in square brackets (e.g., [3]) to distinguish it from other numbers. Multiple citation numbers (each enclosed in brackets) are listed separated by commas (e.g. '([1],[3],[4])').

3.3.3.4 Sequences

A literal sequence of nucleotide bases (e.g., "atgcatt") should be enclosed in quotation marks. A literal sequence may be distinguished from free text, which is also enclosed in double quotation marks, by its context. That is, qualifiers which take free text as their values do not take literal sequences and vice versa.

3.3.4 Qualifier examples

```
Key             Location/Qualifiers
CDS             86..742
                /product="hypoxanthine phosphoribosyltransferase"
                /label=hprt
                /note="hprt catalyzes vital steps in the
                reutilization pathway for purine biosynthesis and
                its deficiency leads to forms of ""gouty""
                arthritis"
rep.origin      234..243
                /direction=left
CDS             109..564
                /usedin=X10009:catalase
```

3.4 Feature labels

The /label= qualifier takes as its value a feature label. Feature labels follow the same naming conventions as other feature table components (e.g., keys and qualifiers). While feature labels are optional, attaching a label to a feature allows it to be referred to unambiguously. Because feature labels have special uses in location descriptors, they are given special attention here.

3.4.1 Purpose

The feature label identifies a feature item within an entry and, when combined with the entry's primary accession number and the name of the database from which it came, is a permanent internationally unique tag for that feature. There are, however, certain situations in which a "permanent" feature may "disappear" from the distributed version of the database and others in which it may be desirable to change a feature's label. See Section 6 for a description of these situations.

3.4.2 Format and conventions

Each feature in a feature table may have a label which must be unique within that entry, but which may be the same as feature labels used in other entries. A feature can be given any label. However, labels containing meaningful abbreviations will be much more easily remembered than non-descriptive labels. Because letter case is not significant, two features within one entry cannot have labels that differ only in case: '16S_rRNA' and '16s_rRNA' could not both be used in the same entry.

The full feature name syntax is as follows:

Database name::primary accession number:feature label

References to a feature should use as much of the full feature name as required to unambiguously identify the feature.

3.4.3 Examples of feature labels

Feature label	Description
adhI	adhI gene coding for alcohol dehydrogenase
tfp35	tail fiber protein 35
3'-ltr	3' long terminal repeat
a1col_x51	prepro-alpha-1-collagen, exon 51
refnum	reference to numbering in a citation
feaA	mutation for strain R45 (labels need not be meaningful)
X10045:diff1	first conflict for the sequence of entry X10045
GB::K10675:catexA	catalase exon A in entry K10675 of the GenBank databank

3.5 Location

3.5.1 Purpose

The location indicates the region of the presented sequence which corresponds to a feature.

3.5.2 Format and conventions

The location contains at least one sequence location descriptor and may contain one or more operators with one or more sequence location descriptors. Base numbers refer to the numbering in the entry. This numbering, which is not necessarily the same as the numbering scheme used in the published report cited, designates the first base (5' end) of the presented sequence as base 1. Base locations beyond the range of the presented sequence may be used (if known) in location descriptors. Bases beyond (before) the 5' end of the presented sequence are numbered starting with the proximal base as 0 and decreasing. Nucleotides beyond the 3' end are numbered beginning with the base number one greater than the presented sequence's length. Location operators and descriptors are discussed in more detail below.

3.5.2.1 Location descriptors

The location descriptor can be one of the following:

(a) a single base number
(b) a site between two indicated base numbers
(c) a single base chosen from within a specified range of bases
(d) the base numbers delimiting a sequence span
(e) a remote entry identifier followed by a local location descriptor (i.e., a-d or g)
(f) a feature label which identifies a region in its own location descriptor
(g) a literal sequence (a string of bases enclosed in quotation marks).

A site between two points (nucleotides), such as endonucleolytic cleavage site, is indicated by listing the two points separated by a carat (^).

A single base chosen from a range or span of bases is indicated by the first base number and the last base number of the range separated by a single period (e.g., '12.21' indicates a single base taken from between the indicated points). The 'less-than' symbol (<) and the 'greater-than' symbol (>) indicate that a range end point is beyond and does not include the specified known base number.

Sequence spans are indicated by the starting base number and the ending base number separated by two periods (e.g., '34..456'). The '<' and '>' symbols may be used with the starting and ending base numbers to indicate that an end point is beyond (and does not include) the specified base number. The starting and ending base positions can be represented as distinct base numbers ('34..456') or as alternatives specified by an operator. A single point chosen from two or more alternative points uses the 'one-of' operator while a single point chosen from a range of points uses the 'x.y' format described above.

Feature labels should be used in location descriptors only when they are required to improve readability. A location in a remote entry (not the entry to which the feature table belongs) can be specified in either of two ways: by specifying the remote entry (by Accession number) followed by a location descriptor which applies to that entry's sequence, or by specifying the remote entry followed by the label of a feature in that entry's feature table.

3.5.2.2 Operators

The location operator is a prefix that specifies what must be done to the indicated sequence to find or construct the location corresponding to the feature. A list of allowable operators is given below with their definitions and most common format.

a) complement(location)
   Find the complement of the presented sequence in the span specified by "location" (i.e., read the complement of the presented strand in its 5'-to-3' direction)

b) join(location,location, ... location)
   The indicated elements should be joined (placed end-to-end) to form one contiguous sequence

c) order(location,location, ... location)
   The elements can be found in the specified order (5' to 3' direction), but nothing is implied about the reasonableness about joining them

d) group(location,location, ... location)
The elements are related and should be grouped together, but no specific order is implied

e) one-of(location,location, ... location)
The specified element can be any one (but only one) of the items listed. This operator may be understood as an exclusive-or (XOR) of its parameters.

f) replace(location,location)
The first indicated location should be replaced by the sequence from the second location; used for insertions, deletions, and variants

Multiple, nested operators are valid where allowed by the context, with the expression being evaluated from right to left.

3.5.3 Location examples

The following is a list of common location descriptors with their meanings:

Location	Description
467	Points to a single base in the presented sequence
340..565	Points to a continuous range of bases bounded by and including the starting and ending bases
<345..500	Indicates that the exact lower boundary point of a feature is unknown. The location begins at some base previous to the first base specified (which need not be contained in the presented sequence) and continues to and includes the ending base
<1..888	The feature starts before the first sequenced base and continues to and includes base 888
102.110	Indicates that the exact location is unknown but that it is one of the bases between bases 102 and 110, inclusive
(23.45)..600	Specifies that the starting point is one of the bases between bases 23 and 45, inclusive, and the end point is base 600
(122.133)..(204.221)	The feature starts at a base between 122 and 133, inclusive, and ends at a base between 204 and 221, inclusive
123^124	Points to a site between bases 123 and 124
145^177	Points to a site between two adjacent bases anywhere between bases 145 and 177
one-of(456,461)..833	Indicates that although the starting point of a feature is unknown, it begins at one of two locations, and continues to and includes the ending base

Location	Description
complement(34..(122.126))	Start at one of the bases complementary to those between 122 and 126 on the presented strand and finish at the base complementary to base 34 (the feature is on the strand complementary to the presented strand)
group((1..100),X00076:prot1)	The sequence from base 1 to base 100 in this entry is related to but not necessarily contiguous with the feature labelled 'prot1' in the entry with primary accession number X00076
join("acct",449..670)	Concatenate the four bases 'acct' to the 5' end of the sequence from bases 449 to 670, inclusive
replace(346,"a")	Replace the base 346 with an 'a'
replace(12..35,"")	Delete the sequence from base 12 to 35, inclusive
replace(800^801,"atttg")	Insert 'atttg' between bases 800 and 801
J00193:hladr	Points to a feature whose location is described in another entry: the feature labelled 'hladr' in the entry (in this database) with primary accession number 'J00193'
J00194:(100..202)	Points to bases 100 to 202, inclusive, in the entry (in this database) with primary accession number 'J00194'

4 Feature table format

This section describes the columnar format used to write this feature table in "flat file" form for distributions of the databases.

4.1 Format example

An example of the feature table format is:

```
FEATURES             Location/Qualifiers
     CDS             100..234
     mutation        replace(115,"a")
                     /phenotype="no activity" /note="Identity of
                     affected protein is unknown" /label=mutx
----+----+----+----+----+----+----+----+----+----+----+----+----+----+
        10        20        30        40        50        60        70
```

Additional examples of the feature table format can be found in Section 5 and Appendix I and in the Annotation Standards Guide.

4.2 Definition of line types

The feature table consists of a header line, which contains the column titles for the table, and the individual feature entries. Each feature entry is composed of a feature descriptor line and qualifier and continuation lines, if needed. The feature descriptor line contains the feature's name, key, and location. If the location cannot be contained on the first line of the feature descriptor, it is continued on a continuation line immediately following the descriptor line. If the feature requires further attributes, feature qualifier lines immediately follow the corresponding feature descriptor line (or its continuation). Qualifier information that cannot be contained on one line continues on the following continuation lines as necessary.

Thus, there are 4 types of feature table lines:

Line type	Content	#/entry	#/feature
Header	Column titles	0 to 1*	N/A
Feature descriptor	Key and location	0 to many*	1
Feature qualifiers	Qualifiers and values	N/A	0 to many
Continuation lines	Feature descriptor or qualifier continuation	0 to many	0 to many

* To allow for entries without feature tables, 0 header and feature descriptor lines are permitted. Any feature table must have one header line and at least one feature descriptor line.

4.3 Data item positions

The position of the data items within the feature descriptor line is as follows:

column position	data item
1-5	blank
6-20	feature key
21	blank
22-80	location

Data on the qualifier and continuation lines begins in column position 22 (the first 21 columns contain blanks). The EMBL format for all lines differs from that described above in that it includes a line type abbreviation in columns 1 and 2.

4.4 Use of blanks

Blanks (spaces) may, in general, be used within the feature location and qualifier values to make the construction more readable. The following rules should be observed:

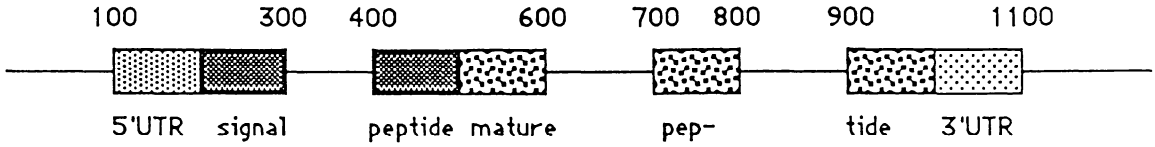
- Names of feature table components may not contain blanks (see Section 3.1)
- Operator names may not be separated from the following open parenthesis (the beginning of the operand list) by blanks.
- Qualifiers may not be separated from the preceding slash or the following equals sign (if one) by blanks

5 Examples of sequence annotation

Note that the examples given below are only samples of one way a sequence may be annotated and other ways may also be acceptable.

5.1 Protein-coding region

Example: Prototypical eukaryotic gene




```
FEATURES             Location/Qualifiers
     5'UTR           100..200
     exon            100..300
                     /number=1
     intron          301..400
                     /number=1
     exon            401..600
                     /number=2
     intron          601..700
                     /number=2
     exon            701..800
                     /number=3
     intron          801..900
                     /number=3
     3'UTR           1002..1100
     exon            901..1100
                     /number=4
     sig_peptide     join(201..300,401..501)
     mat_peptide     join(502..600,701..800,901..1001)
                     /product="prototypical protein"
     CDS             join(201..300,401..600,701..800,901..1001)
                     /product="prototypical protein"
     mRNA            join(100..300,401..600,701..800,901..1100)
     prim_transcript 100..1100
```

5.2 Structural RNA

Example: Bacillus Q 5S rRNA precursor (X01353)

```
FEATURES             Location/Qualifiers
     5'clip          1..21
                     /label=5s-start
     rRNA            22..137
                     /label=5srna
     3'clip          138..172
                     /label=5s-end
     pre-5srna       prim_transcript join(5s-start,5srna,5s-end)
                     /product="5S ribosomal RNA precursor"
```

5.3 Regulatory regions

Example: Human opsin gene (K02281)

```
FEATURES             Location/Qualifiers
     CAAT.signal     122..127
                     /label=ops-cat
     TATA.signal     171..177
                     /label=ops-tata
     5'UTR           200..294
                     /note="opsin transcript 1"
     5'UTR           202..294
                     /note="opsin transcript 2"
     exon            200..655
                     /note="opsin transcript 1" /number=1
     exon            202..655
                     /note="opsin transcript 2" /number=1
     intron          656..2438
     exon            2439..2607
                     /number=2
     intron          2608..3812
     exon            3813..3978
                     /number=3
     intron          3979..4094
     exon            4095..4334
                     /number=4
     intron          4335..5167
     3'UTR           5279..>5279
     exon            5168..>5279
                     /number=5
     polyA.signal    5642..5647
     polyA.signal    6698..6903
     mRNA            join(200..655,2439..2607,3813..3978,4095..4334,
                     5168..>5279) /note="opsin transcript 1"
     mRNA            join(202..655,2439..2607,3813..3978,4095..4334,
                     5168..>5279) /note="opsin transcript 2"
     CDS             join(295..655,2439..2607,3813..3978,4095..4334,
                     5168..5278) /product="opsin"
     mat.peptide     join(295..655,2439..2607,3813..3978,4095..4334,
                     5168..5275) /product="opsin"
     promoter        order(ops-cat,ops-tata)
```

6 Known limitations of this feature table design

During the development of this feature table design numerous choices between simplicity and representational power had to be made. In order to create a design which was capable of representing the most common features of biological significance, a certain degree of complexity in the syntax was guaranteed. However, to limit that level of complexity, certain limitations of the design syntax have been accepted. This means that some rare and complicated relationships among features will be inadequately represented by this syntax. In addition, in order to make a rapid conversion of the existing feature tables into the new format, some implementation limitations are also expected. Chief among these are:

- Merging entries can result in changed individual feature labels and (of course) a loss of the full entry name -- the constituent entries and features will be maintained in the permanent copies of the data banks, but a given database distribution (view) may contain only the merged entries. Within that view, it will appear that 'permanent' features labels have disappeared.

- The current syntax limits annotation to features of the presented sequence. This feature table design does allow annotation of variants of the presented sequence through the use of change features (such as mutation, variation, etc.). In general, however, features of such sequence variants cannot be annotated as features of the variant only. That is, while a feature which exists solely in one of the annotated sequence variants (an altered gene product in a mutant, for example) can be annotated in the feature table, that it is a feature of the variant alone can be indicated only in a textual note. When an entry contains change or difference features, consideration should be given to the dependence of other features' existence upon the changes. In these cases, caution should be exercised when analyzing or extracting sequence features.

- Many data items from current entries which should be expressed with features qualifiers will be maintained as free text notes indefinitely. All new entries will be annotated according to the format and conventions described in this document and its appendices and (in more detail) in the Feature Table Annotation Standards Guide.

- This feature table syntax allows feature labels to be used as elements of a location descriptor without providing a mechanism to ensure that such references can be resolved to a location descriptor using only base-positions and literal sequences. It is the responsibility of feature table creators and maintainers to ensure that all location descriptor references can be resolved.

- The present syntax gives several legal options for expressing certain feature locations. The most obvious of these is that a sequence span can be expressed either by designating the base numbers of the range or by naming another feature which defines the intended span (and has been given a label with the /label= qualifier). Although the distributed versions of the databases will resolve most feature locations to numeric base range specifications, it may not be desirable or possible to do so in all cases. Robust software must be capable of parsing all legal feature location syntaxes, including those which require resolving feature location references to feature labels in the same feature table or in other entries.

- Feature labels will only be changed for good reason; however, some will occasionally be changed. Within a given release of a database, all references will be updated to the new name, but users should be aware that from one release to the next it may be necessary to consult a table (maintained at each database site) of the changed feature labels. Likewise references across databases may not be resolvable during an interim period while remote features' references are being updated.

7 Appendices

7.1 Appendix I EMBL and GenBank entries

EMBL Format

```
ID   ACAC01    standard; DNA; 1571 BP.
XX
AC   V00002;
XX
DT   26-JAN-1984  (species spelling)
DT   15-JUN-1983  (first entry)
XX
DE   Acanthamoeba castelani gene encoding actin I.
XX
KW   actin.
XX
OS   Acanthamoeba castellani (amoeba, amibe, Amoebe)
OC   Eukaryota; Protozoa; Rhizopoda.
XX
RN   [1]   (bases 1-1571; enum. 1 to 1571)
RA   Nellen W., Gallwitz D.;
RT   "Actin genes and actin messenger RNA in Acanthamoeba
RT   castellani. Nucleotide sequence of the split actin
RT   gene I";
RL   J. Mol. Biol. 159:1-18(1982).
XX
FH   Key             Location/Qualifiers
FH
FT   prim_transcript one-of(90,91)..>1443
FT   exon            one-of(90,91)..451
                     /number=1
FT   intron          452..580
                     /number=1
FT   exon            581..>1443
                     /number=2
FT   CDS             join((137..451),(581..1390))
                     /product="actin I"
XX
SQ   Sequence  1571 BP;  313 A;  535 C;  389 G;  334 T;
----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+
        10        20        30        40        50        60        70
```

GenBank Format

```
LOCUS       HUMPALB       614 bp ss-mRNA            PRI       15-JUN-1988
DEFINITION  Human prealbumin mRNA, complete cds.
ACCESSION   M10605
KEYWORDS    prealbumin.
SOURCE      Human liver, cDNA to mRNA, clone PA7.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Vertebrata; Tetrapoda; Mammalia;
            Eutheria; Primates; Anthropoidea; Hominoidea; Hominidae.
REFERENCE   1  (bases 1 to 614)
  AUTHORS   Wallace,M.R., Naylor,S.L., Kluve-Beckerman,B., Long,G.L.,
            McDonald,L., Shows,T.B. and Benson,M.D.
  TITLE     Localization of the human prealbumin gene to chromosome 18
  JOURNAL   Biochem Biophys Res Commun 129, 753-758 (1985)
  STANDARD  simple staff_review
COMMENT     Draft entry and sequence in computer readable form for [1]
kindly
            provided by M.R.Wallace, 26-DEC-1985.
FEATURES             LOCATION/QUALIFIERS
     sig_peptide     26..85
     mat_peptide     86..466
                     /product="prealbumin"
     CDS             26..469
                     /product="prealbumin"
     mRNA            <1..614
BASE COUNT      148 a    162 c    155 g    149 t
ORIGIN      247 bp upstream of AluI site; chromosome 18.
//
----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+
        10        20        30        40        50        60        70
```

7.2 Appendix II Feature table: Backus-Naur form

Feature table is an optional part of an entry. Full entry syntax is specified elsewhere.

feature_table ::= <feature_table_header><feature_table_body>

feature_table_header ::= `FH  Key  Location/Qualifiers` |
`FEATURES  Location/Qualifiers`

feature_table_body ::= <feature> | <feature_table_body><feature>

At least one feature is required.

feature ::= <feature_key><feature_details>

Key is required, location required, qualifier list optional

feature_key ::= <symbol> | -

feature_details ::= <location><qualifier_list> | <location>

There exists a table of legal keys. "-" is a placeholder for no key.

location ::= <absolute_location> | <feature_name> | <literal_sequence> |
<functional_operator>(<location_list>)

absolute_location ::= <local_location> | <path> : <local_location>

path ::= <database> :: <primary_accession> | <primary_accession>

feature_name ::= <path>:<feature_label> | <feature_label>

feature_label :== <symbol>

local_location ::= <base_position> | <between_position> | <base_range>

location_list ::= <location> | <location_list>,<location>

functional_operator ::= <symbol>

base_position ::= <integer> | <low_base_bound> | <high_base_bound> | <two_base_bound>

low_base_bound ::= > <integer>

high_base_bound ::= < <integer>

two_base_bound ::= <base_position>.<base_position>

between_position ::= <base_position>^<base_position>

base_range ::= <base_position>..<base_position>

database ::= <symbol>

primary_accession ::= <symbol>

literal_sequence ::= <sequence_character> | <literal_sequence><sequence_character>

sequence_character ::= a | b | c | d | g | h | k | m | n | r | s | t | u | v | w | y

qualifier_list ::= <qualifier> | <qualifier_list><qualifier>

qualifier ::= /<qualifier_name> | /<qualifier_name>=<value>

qualifier_name ::= <symbol>

value ::= <simple_value> | (<value_list>) | (<tagged_value_list>)

simple_value ::= <integer> | <location> | <reference_number> | "<text_string>" | <symbol>

value_list ::= <value> | <value_list>,<value>

tagged_value_list ::= <tagged_value> | <tagged_value_list>,<tagged_value>

tagged_value ::= <tag>:<value>

tag ::= <symbol>

reference_number ::= [<unsigned_integer>]

symbol ::= <letter> | <symbol><symbol_character> | <symbol_character><symbol>

text_string ::= <string_character>| <text_string><string_character>

unsigned_integer ::= <digit> | <unsigned_integer><digit>

integer ::= <unsigned_integer> | - <unsigned_integer>

string_character ::= <letter> | <digit> | <punctuation> | ""

symbol_character ::= <up_case_letter> | <low_case_letter> |<digit> | _ | - | ' | *

letter ::= <up_case_letter> | <low_case_letter>

up_case_letter ::= A | B| ... | Z

low_case_letter ::= a | b | ... | z

digit ::= 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

punctuation ::= <space> | ! | # | $ | % | & | ' | (|) | * | + | , | - | . | / | : | ; | < | = | > | ? | @ | [
| \ |] | ^ | _ | ` | { | <bar> | } | ~

bar ::= |

space ::= ascii 32

7.3 Appendix III Feature keys reference

7.3.1 Feature key relationship tree

A. misc_feature

1. misc_difference
   a) conflict
   b) unsure
   c) old_sequence
   d) mutation
   e) variation
   f) allele
   g) modified_base

2. misc_signal
   a) promoter
      1) CAAT_signal
      2) TATA_signal
      3) -35_signal
      4) -10_signal
      5) GC_signal
   b) RBS
   c) polyA_signal
   d) enhancer
   e) attenuator
   f) terminator
   g) rep_origin

3. misc_RNA
   a) prim_transcript
      1) precursor_RNA
         a)mRNA
         b) 5'clip
         c) 3'clip
         d) 5'UTR
         e) 3'UTR
         f) exon
         g) CDS
            1) sig_peptide
            2) transit_peptide
            3) mat_peptide
         h) intron
         i) polyA_site
         j) rRNA
         k) tRNA
         l) scRNA
         m) snRNA

4. repeat_region
   a) repeat_unit
   b) LTR
   c) satellite

5. misc_binding
   1) primer_bind
   2) protein_bind

6. misc_recomb
   a) cellular
   b) iDNA
   c) insertion_seq
   d) transposon
   e) provirus
   f) virion

7. misc_structure
   a) stem_loop
   b) D-loop

7.3.2 Feature key reference manual

The following manual has been organized according to the following format:

Feature Key	the feature key name
Definition	the definition of the key
Mandatory qualifiers	qualifiers required with the key; if there are no mandatory qualifiers, this field is omitted.
Optional qualifiers	optional qualifiers associated with the key
Organism scope	valid organisms for the key; if the scope is any organism, this field is omitted.
Molecule scope	valid molecule types; if the scope is any molecule type, this field is omitted.
Old GB key	the old related GenBank key(s); if there was no corresponding feature key in the old GenBank feature table, the field is omitted.
Old EMBL key	the old related EMBL key(s); if there was no corresponding feature key in the old EMBL feature table, the field is omitted.
References	citations of published reports, usually supporting the feature consensus sequence
Comment	comments and clarifications

Abbreviations:

accnum	an entry primary accession number
<amino_acid>	abbreviation for amino acid
<base_range>	location descriptor for a simple range of bases
<bool>	Boolean truth value. Valid values are **yes** and **no**
<evidence_value>	value indicating the nature of supporting evidence. **experimental** is the only currently valid value.
feature_label	the feature label (follows naming conventions for all feature table components)
<integer>	unsigned integer value
<location>	general feature location descriptor
<modified_base>	abbreviation for modified nucleoside base
[number]	integer representing number of citation in entry's reference list
<repeat_type>	value indicating the organization of a repeated sequence. Currently valid values are **tandem, inverted, flanking, terminal, direct, dispersed,** and **other**
"text"	any text or character string

Feature Key	**allele**
Definition	a related individual or strain contains stable, alternative forms of the same gene which differs from the presented sequence at this location (and perhaps others)
Optional qualifiers	/citation=[number] /frequency="text" /gene="text" /label=feature_label /note="text" /phenotype="text" /product="text" /standard_name="text" /type="text" /usedin=accnum:feature_label
Old GB key	allele
Old EMBL key	ALLELE
Comment	format of location is: replace(seq_location,variation) where variation is the sequence difference. The /type qualifier is used to specify a strain name.

Feature Key	**attenuator**
Definition	1) region of DNA at which regulation of termination of transcription occurs, which controls the expression of some bacterial operons; 2) sequence segment located between the promoter and the first structural gene that causes partial termination of transcription
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /label=feature_label /note="text" /partial /phenotype="text" /usedin=accnum:feature_label
Organism scope	prokaryotes
Molecule scope	DNA
Old EMBL key	ATTEN

Feature Key	**CAAT_signal**
Definition	CAAT box; part of a conserved sequence located about 75 bp upstream of the start point of eukaryotic transcription units which may be involved in RNA polymerase binding; consensus=GG(C or T)CAATCT [1,2]

Optional qualifiers	/citation=[number] /evidence=<evidence_value> /label=feature_label /note="text" /usedin=accnum:feature_label
Organism scope	Eukaryotes and eukaryotic viruses
Molecule scope	DNA
Old EMBL key	PRM
References	[1] Efstratiadis, A. et al. Cell **21**, 653-668 (1980) [2] Nevins, J.R. "The pathway of eukaryotic mRNA formation" Ann Rev Biochem **52**, 441-466 (1983)

Feature Key **CDS**

Definition	coding sequence; sequence of nucleotides that corresponds with the sequence of amino acids in a protein (location includes stop codon)
Optional qualifiers	/citation=[number] /codon=(seq:"text",aa:<amino_acid>) /codon_start=<location> /EC_number="text" /evidence=<evidence_value> /function="text" /gene="text" /label=feature_label /note="text" /number=<integer> /organism="text" /partial /product="text" /pseudo /standard_name="text" /transl_except=(pos:<base_range>,aa:<amino_acid>) /usedin=accnum:feature_label
Old GB key	pept
Old EMBL key	CDS
Comment	/codon_start has valid value of the single base location of a codon start; /codon is used to specify unusual genetic codes, including rare-usage start codons, organellar codes, etc.; /codon is used to describe a single codon exception to the code defined as the "normal" code for the organism; it implies that the translation it specifies is used throughout the feature /codon specifies that a specific codon specified by "seq" codes for the amino acid or stop codon specified by "aa";

/transl_except is used to specify a single codon the translation of which does not conform to the genetic code defined by the "normal" code for organism or by the /codon qualifiers given for the feature

Feature Key	**cellular**
Definition	cellular genomic sequences that have recombined with a foreign sequence (from the organism specified in /organism qualifier)
Mandatory qualifiers	/organism="text"
Optional qualifiers	/citation=[number] /label=feature_label /note="text" /usedin=accnum:feature_label
Molecule scope	DNA
Old GB key	cell
Old EMBL key	CELL
Comment	the foreign recombining sequence is typically retroviral, but may be an insertion sequence, transposon, DNA in a chimeric engineered sequence, or other

Feature Key	**conflict**
Definition	independent determinations of the "same" sequence differ at this site or region
Mandatory qualifiers	/citation=[number]
Optional qualifiers	/note="text" /usedin=accnum:feature_label
Old GB key	conflict
Old EMBL key	CONFLICT; ERROR
Comment	format of location is: replace(seq_location,conflict_location) where conflict_location is the sequence difference

Feature Key	**D-loop**
Definition	displacement loop; a region within mitochondrial DNA in which a short stretch of RNA is paired with one strand of DNA, displacing the original partner DNA strand in this region; also used to describe the displacement of a region of one strand of duplex DNA by a single stranded invader in the reaction catalyzed by RecA protein

Optional qualifiers
/citation=[number]
/label=feature_label
/note="text"
/partial
/usedin=accnum:feature_label

Molecule scope DNA

Old GB key D-loop

Feature Key enhancer

Definition a cis-acting sequence that increases the utilization of (some) eukaryotic promoters, and can function in either orientation and in any location (upstream or downstream) relative to the promoter

Optional qualifiers
/citation=[number]
/evidence=<evidence_value>
/label=feature_label
/note="text"
/partial
/standard_name="text"
/usedin=accnum:feature_label

Organism scope Eukaryotes and eukaryotic viruses

Feature Key exon

Definition region of genome that codes for portion of spliced mRNA; may contain 5'UTR, all CDSs, and 3' UTR

Optional qualifiers
/citation=[number]
/codon=(seq:"text",aa:<amino_acid>)
/codon_start=<location>
/EC_number="text"
/evidence=<evidence_value>
/function="text"
/gene="text"
/label=feature_label
/note="text"
/number=<integer>
/organism="text"
/partial
/product="text"
/pseudo
/standard_name="text"
/transl_except=(pos:<base_range>,aa:<amino_acid>)
/usedin=accnum:feature_label

Old GB key pept

Old EMBL key CDS

Comment	/codon_start has valid value of the single base location of a codon start; /codon is used to specify unusual genetic codes, including rare-usage start codons, organellar codes, etc.; /codon is used to describe a single codon exception to the code defined as the "normal" code for the organism; it implies that the translation it specifies is used throughout the feature /codon specifies that a specific codon specified by "seq" codes for the amino acid or stop codon specified by "aa"; /transl_except is used to specify a single codon the translation of which does not conform to the genetic code defined by the "normal" code for organism or by the /codon qualifiers given for the feature

Feature Key **GC_signal**

Definition	GC box; a conserved GC-rich region located upstream of the start point of eukaryotic transcription units which may occur in multiple copies or in either orientation; consensus=GGGCGG
Optional qualifiers	/citation=[number] /label=feature_label /note="text" /usedin=accnum:feature_label
Organism scope	Eukaryotes and eukaryotic viruses
Old EMBL key	PRM

Feature Key **iDNA**

Definition	intervening DNA; DNA which is eliminated through any of several kinds of recombination
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /function="text" /label=feature_label /note="text" number=<integer> /partial /standard_name="text" /usedin=accnum:feature_label
Molecule scope	DNA
Old GB key	iDNA
Old EMBL key	IVS
Comment	e.g., in the somatic processing of immunoglobulin genes.

Feature Key	insertion_seq
Definition	insertion sequence; IS; a small transposon that carries only the genes needed for its own transposition
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /function="text" /label=feature_label /note="text" /organism="text" /partial /standard_name="text" /usedin=accnum:feature_label
Organism scope	Eukaryotes, Prokaryotes
Molecule scope	DNA
Old GB key	trns
Old EMBL key	INSSQ
Comment	the name of the insertion sequence may be given with the /standard_name qualifier; can /organism be used for the normal host?

Feature Key	intron
Definition	a segment of DNA that is transcribed, but removed from within the transcript by splicing together the sequences (exons) on either side of it
Optional qualifiers	/citation=[number] /cons_splice=(5'site:<bool>,3'site:<bool>) /evidence=<evidence_value> /function="text" /label=feature_label /note="text" /number=<integer> /partial /standard_name="text" /usedin=accnum:feature_label
Old GB key	IVS
Old EMBL key	IVS
Comment	cons_splice is used only when one of the intron's splice sites does not match the GT...AG consensus. Reference for this consensus?

Feature Key	**LTR**
Definition	long terminal repeat, a sequence directly repeated at both ends of a defined sequence, of the sort typically found in retroviruses
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /function="text" /label=feature_label /note="text" /partial /standard_name="text" /usedin=accnum:feature_label
Old GB key	LTR
Old EMBL key	RPT

Feature Key	**mat_peptide**
Definition	mature peptide or protein coding sequence; coding sequence for the mature or final peptide or protein product following post-translational modification. the location does not include the stop codon (unlike the corresponding CDS)
Optional qualifiers	/citation=[number] /codon=(seq:"text",aa:<amino_acid>) /codon_start=<location> /EC_number="text" /evidence=<evidence_value> /function="text" /gene="text" /label=feature_label /note="text" /organism="text" /partial /product="text" /pseudo /standard_name="text" /transl_except=(pos:<base_range>,aa:<amino_acid>) /usedin=accnum:feature_label
Old GB key	matp; pept
Old EMBL key	CDS
Comment	/codon_start has valid value of the single base location of a codon start; /codon is used to specify unusual genetic codes, including rare-usage start codons, organellar codes, etc.; /codon is used to describe a single codon exception to the code defined as the "normal" code for the organism; it implies that the translation it specifies is used throughout the feature

/codon specifies that a specific codon specified by "seq" codes for the amino acid or stop codon specified by "aa";
/transl_except is used to specify a single codon the translation of which does not conform to the genetic code defined by the "normal" code for organism or by the /codon qualifiers given for the feature

Feature Key	misc_binding
Definition	site in nucleic acid which covalently or non-covalently binds another moiety that cannot be described by any other Binding key (primer_bind or protein_bind)
Mandatory qualifiers	/bound_moiety="text"
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /function="text" /label=feature_label /note="text" /usedin=accnum:feature_label
Old GB key	binding
Comment	note that the key RBS is used for ribosome binding sites

Feature Key	misc_difference
Definition	feature sequence is different from that presented in the entry and cannot be described by any other Difference key (conflict, unsure, old_sequence, mutation, variation, allele, or modified_base)
Optional qualifiers	/citation=[number] /label=feature_label /note="text" /standard_name="text" /usedin=accnum:feature_label
Old GB key	attack; cutds; cutss
Comment	format of location is: replace(seq_location,change_location) where change_location is a sequence descriptor or feature name that replaces the indicated sequence to yield the difference sequence

Feature Key	misc_feature
Definition	region of biological interest which cannot be described by any other feature key; a new or rare feature
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /function="text"

/label=feature_label
/note="text"
/number=<integer>
/phenotype="text"
/product="text"
/pseudo
/standard_name="text"
/usedin=accnum:feature_label

Old GB key: site

Old EMBL key: SITE

Comment: To be invoked infrequently. This key should not be used when the need is merely to mark a region in order to comment on it or to use it in another feature's location; use the '-' pseudo-key instead.

Feature Key **misc_recomb**

Definition: site of any generalized, site-specific or replicative recombination event where there is a breakage and reunion of duplex DNA that cannot be described by other Recombination keys (cellular, iDNA, insertion_seq, transposon, provirus, and virion)

Mandatory qualifiers: /organism="text"

Optional qualifiers:
/citation=[number]
/evidence=<evidence_value>
/label=feature_label
/note="text"
/standard_name="text"
/usedin=accnum:feature_label

Molecule scope: DNA

Old GB key: recomb

Old EMBL key: SOMRECOMB

Feature Key **misc_RNA**

Definition: any transcript or RNA product that cannot be defined by other RNA keys (prim_transcript, precursor_RNA, mRNA, 5'clip, 3'clip, 5'UTR, 3'UTR, exon, CDS, sig_peptide, transit_peptide, mat_peptide, intron, polyA_site, rRNA, tRNA, scRNA, and snRNA)

Optional qualifiers:
/citation=[number]
/evidence=<evidence_value>
/function="text"
/gene="text"
/label=feature_label
/note="text"

/partial
/product="text"
/pseudo
/standard_name="text"
/usedin=accnum:feature_label

Old GB key RNA; iRNA

Feature Key **misc_signal**

Definition any region containing a signal controlling or altering gene function or expression that cannot be described by other Signal keys (promoter, CAAT_signal, TATA_signal, -35_signal, -10_signal, GC_signal, RBS, polyA_signal, enhancer, attenuator, terminator, and rep_origin)

Optional qualifiers /citation=[number]
/evidence=<evidence_value>
/function="text"
/label=feature_label
/note="text"
/partial
/phenotype="text"
/standard_name="text"
/usedin=accnum:feature_label

Old GB key signal

Feature Key **misc_structure**

Definition any secondary or tertiary structure or conformation that cannot be described by other Structure keys (stem_loop and D-loop)

Optional qualifiers /citation=[number]
/evidence=<evidence_value>
/function="text"
/label=feature_label
/note="text"
/partial
/standard_name="text"
/usedin=accnum:feature_label

Feature Key **modified_base**

Definition the indicated nucleotide is a modified nucleotide and should be substituted for by the indicated molecule (given in the mod_base qualifier value)

Mandatory qualifiers /mod_base=<modified_base>

Optional qualifiers /citation=[number]
/evidence=<evidence_value>

/frequency="text"
/label=feature_label
/note="text"
/usedin=accnum:feature_label

Old GB key — modified; methyl

Old EMBL key — MODBASE

Comment — the mod_base value is limited to the restricted vocabulary for modified base abbreviations

Feature Key — **mRNA**

Definition — messenger RNA; includes 5'untranslated region (5'UTR), coding sequences (CDS, exon) and 3'untranslated region (3'UTR)

Optional qualifiers
/citation=[number]
/evidence=<evidence_value>
/function="text"
/gene="text"
/label=feature_label
/note="text"
/partial
/standard_name="text"
/usedin=accnum:feature_label

Old GB key — mRNA; pept

Old EMBL key — MSG

Feature Key — **mutation**

Definition — a related strain has an abrupt, inheritable change in the sequence at this location

Optional qualifiers
/citation=[number]
/frequency="text"
/gene="text"
/label=feature_label
/note="text"
/phenotype="text"
/product="text"
/standard_name="text"
/type="text"
/usedin=accnum:feature_label

Old GB key — mut

Old EMBL key — MUTANT

Comment	format of location is: replace(seq_location,mutation_location) where mutation_location is the sequence difference and seq_location specifies the location in the presented sequence to be replaced. The /type qualifier may be used to specify a strain name.

Feature Key **old_sequence**

Definition	the presented sequence revises a previous version of the sequence at this location
Mandatory qualifiers	/citation=[number]
Optional qualifiers	/note="text" /usedin=accnum:feature_label
Old GB key	revision
Old EMBL key	REVISION
Comment	format of location is: replace(seq_location,change_loc) where change_loc represents the old sequence (before revision) and is usually a literal sequence and seq_location is the location in the presented (revised) sequence at which the revision took place. To regenerate the sequence as it was before revision, replace the seq_location portion of the presented sequence with the sequence specified by change_loc

Feature Key **polyA_signal**

Definition	recognition region necessary for endonuclease cleavage of an RNA transcript that is followed by polyadenylation; consensus=AATAAA [1]
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /label=feature_label /note="text" /usedin=accnum:feature_label
Organism scope	Eukaryotes and eukaryotic viruses
Old EMBL key	SITE
References	[1] Proudfoot, N. and Brownlee, G.G. Nature **263**, 211-214 (1976)
Comment	this key is NOT indicating the actual site of poly-adenylation and therefore does NOT replace EMBL's POLYA.

Feature Key **polyA_site**

Definition	site on an RNA transcript to which will be added adenine residues by post-transcriptional polyadenylation

Optional qualifiers /citation=[number]
/evidence=<evidence_value>
/label=feature_label
/note="text"
/usedin=accnum:feature_label

Organism scope Eukaryotes and eukaryotic viruses

Old EMBL key POLYA

Feature Key **precursor_RNA**

Definition any RNA species that is not yet the mature RNA product; may include 5' clipped region (5'clip), 5' untranslated region (5'UTR), coding sequences (CDS, exon), intervening sequences (intron), 3' untranslated region (3'UTR), and 3' clipped region (3'clip)

Optional qualifiers /citation=[number]
/evidence=<evidence_value>
/function="text"
/gene="text"
/label=feature_label
/note="text"
/partial
/standard_name="text"
/usedin=accnum:feature_label

Old EMBL key TRANSCR

Comment this key is used for any RNA which may be the result of some post-transcriptional processing. If the RNA in question is known not to have been processed, used the prim_transcript key.

Feature Key **prim_transcript**

Definition primary (initial, unprocessed) transcript; includes 5' clipped region (5'clip), 5' untranslated region (5'UTR), coding sequences (CDS, exon), intervening sequences (intron), 3' untranslated region (3'UTR), and 3' clipped region (3'clip)

Optional qualifiers /citation=[number]
/evidence=<evidence_value>
/function="text"
/gene="text"
/label=feature_label
/note="text"
/partial
/standard_name="text"
/usedin=accnum:feature_label

Old EMBL key TRANSCR

Feature Key	**primer_bind**
Definition	non-covalent primer binding site for initiation of replication, transcription, or reverse transcription
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /label=feature_label /note="text" /standard_name="text" /usedin=accnum:feature_label
Old GB key	binding

Feature Key	**promoter**
Definition	region on a DNA molecule involved in RNA polymerase binding to initiate transcription
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /function="text" /label=feature_label /note="text" /partial /phenotype="text" /pseudo /standard_name="text" /usedin=accnum:feature_label
Molecule scope	DNA
Old EMBL key	PRM

Feature Key	**protein_bind**
Definition	non-covalent protein binding site on nucleic acid
Mandatory qualifiers	/bound_moiety="text"
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /function="text" /label=feature_label /note="text" /standard_name="text" /usedin=accnum:feature_label
Comment	note that RBS is used for ribosome binding sites

Feature Key	provirus
Definition	proviral sequence specified by /organism
Mandatory qualifiers	/organism="text"
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /label=feature_label /note="text" /partial /usedin=accnum:feature_label
Organism scope	Eukaryotes, Prokaryotes
Molecule scope	DNA
Old GB key	prov
Old EMBL key	PROVRL
Comment	The value of the /organism qualifier is the name of the virus. The host organism (that into the genome of which the provirus has inserted) is identified in the Organism field of the entry.

Feature Key	RBS
Definition	ribosome binding site
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /label=feature_label /note="text" /pseudo /standard_name="text" /usedin=accnum:feature_label
Old EMBL key	RBS
References	[1] Shine, J. and Dalgarno, L. Proc Natl Acad Sci USA 71, 1342-1346 (1974) [2] Gold, L. et al. Ann Rev Microb 35, 365-403 (1981)
Comment	in prokaryotes, known as the Shine-Dalgarno sequence: is located 5 to 9 bases upstream of the initiation codon; consensus GGAGGT [1,2]

Feature Key	repeat_region
Definition	region of genome containing repeating units
Optional qualifiers	/citation=[number] /evidence=<evidence_value>

/function="text"
/label=feature_label
/note="text"
/partial
/rpt_type=<repeat_type>
/rpt_family="text"
/rpt_unit=feature_label
/standard_name="text"
/usedin=accnum:feature_label

Old GB key rpt

Old EMBL key RPT

Feature Key **repeat_unit**

Definition single repeat element

Optional qualifiers /citation=[number]
/evidence=<evidence_value>
/function="text"
/label=feature_label
/note="text"
/partial
/rpt_family="text"
/rpt_type=<repeat_type>
/usedin=accnum:feature_label

Old GB key rpt

Old EMBL key RPT

Comment preferred usage is to annotate the /rpt_family and rpt_type qualifiers on the repeat_region, not on the repeat_unit(s)

Feature Key **rep_origin**

Definition origin of replication; starting site for duplication of nucleic acid to give two identical copies

Optional qualifiers /citation=[number]
/direction=value
/evidence=<evidence_value>
/label=feature_label
/note="text"
/partial
/standard_name="text"
/usedin=accnum:feature_label

Old GB key orgrpl

Old EMBL key ORGRPL

Comment	/direction has valid values: RIGHT, LEFT, or BOTH

Feature Key **rRNA**

Definition	mature ribosomal RNA ; the RNA component of the ribonucleoprotein particle (ribosome) which assembles amino acids into proteins
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /function="text" /gene="text" /label=feature_label /note="text" /partial /product="text" /pseudo /standard_name="text" /usedin=accnum:feature_label
Old GB key	rRNA
Old EMBL key	RRNA
Comment	rRNA sizes should be annotated with the /product qualifier. It may be desirable to define a restricted vocabulary for ribosomal RNA size classes.

Feature Key **satellite**

Definition	many tandem repeats (identical or related) of a short basic repeating unit; many have a base composition or other property different from the genome average that allows them to be separated from the bulk (main band) genomic DNA
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /label=feature_label /note="text" /partial /rpt_type=<repeat_type> /rpt_family="text" /standard_name="text" /usedin=accnum:feature_label
Molecule scope	DNA
Old GB key	rpt
Old EMBL key	RPT

Comment — use the satellite key to identify the entire region of satellite sequence within an entry; use repeat_unit to identify individual repeated units (one is generally sufficient) of the satellite; what does the use of /partial imply?

Feature Key — **scRNA**

Definition — small cytoplasmic RNA; any one of several small cytoplasmic RNA molecules present in the cytoplasm and (sometimes) nucleus of a eukaryote

Optional qualifiers
/citation=[number]
/evidence=<evidence_value>
/function="text"
/gene="text"
/label=feature_label
/note="text"
/partial
/product="text"
/pseudo
/standard_name="text"
/usedin=accnum:feature_label

Old EMBL key — SITE; SCRNA

Feature Key — **sig_peptide**

Definition — signal peptide coding sequence; coding sequence for an N-terminal domain of a secreted protein; this domain is involved in attaching nascent polypeptide to the membrane; leader sequence

Optional qualifiers
/citation=[number]
/codon=(seq:"text",aa:<amino_acid>)
/codon_start=<location>
/evidence=<evidence_value>
/function="text"
/label=feature_label
/note="text"
/organism="text"
/partial
/product="text"
/pseudo
/standard_name="text"
/transl_except=(pos:<base_range>,aa:<amino_acid>)
/usedin=accnum:feature_label

Old GB key — sigp; pept

Old EMBL key — CDS

Comment — /codon_start has valid value of the single base location of a codon start;

/codon is used to specify unusual genetic codes, including rare-usage start codons, organellar codes, etc.;
/codon is used to describe a single codon exception to the code defined as the "normal" code for the organism; it implies that the translation it specifies is used throughout the feature
/codon specifies that a specific codon specified by "seq" codes for the amino acid or stop codon specified by "aa";
/transl_except is used to specify a single codon the translation of which does not conform to the genetic code defined by the "normal" code for organism or by the /codon qualifiers given for the feature

Feature Key	snRNA
Definition	small nuclear RNA; any one of many small RNA species confined to the nucleus; several of the snRNAs are involved in splicing or other RNA processing reactions
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /function="text" /gene="text" /label=feature_label /note="text" /partial /product="text" /pseudo /standard_name="text" /usedin=accnum:feature_label
Old GB key	uRNA
Old EMBL key	SITE; URNA; SNRNA

Feature Key	stem_loop
Definition	hairpin; a double-helical region formed by base-pairing between adjacent (inverted) complementary sequences in a single strand of RNA or DNA
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /function="text" /label=feature_label /note="text" /partial /standard_name="text" /usedin=accnum:feature_label
Comment	see the Feature Table Annotation Standards Guide for the annotation convention for stem loop regions

Feature Key	**TATA_signal**
Definition	TATA box; Goldberg-Hogness box; a conserved AT-rich septamer found about 25 bp before the start point of each eukaryotic RNA polymerase II transcript unit which may be involved in positioning the enzyme for correct initiation; consensus=TATA(A or T)A(A or T) [1,2]
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /label=feature_label /note="text" /usedin=accnum:feature_label
Organism scope	Eukaryotes and eukaryotic viruses
Molecule scope	DNA
Old EMBL key	PRM
References	[1] Efstratiadis, A. et al. Cell **21**, 653-668 (1980) [2] Corden, J., et al. "Promoter sequences of eukaryotic protein-encoding genes" Science **209**, 1406-1414 (1980)

Feature Key	**terminator**
Definition	sequence of DNA located either at the end of the transcript or adjacent to a promoter region that causes RNA polymerase to terminate transcription; may also be site of binding of repressor protein
Optional qualifiers	/citation=[number] /evidence=<evidence_value> /label=feature_label /note="text" /partial /standard_name="text" /usedin=accnum:feature_label
Molecule scope	DNA
Old EMBL key	OPR

Feature Key	**transit_peptide**
Definition	transit peptide coding sequence; coding sequence for an N-terminal domain of a nuclear-encoded organellar protein; this domain is involved in post-translational import of the protein into the organelle
Optional qualifiers	/citation=[number] /codon=(seq:"text",aa:<amino_acid>) /codon_start=<location> /evidence=<evidence_value> /function="text"

/label=feature_label
/note="text"
/organism="text"
/partial
/product="text"
/standard_name="text"
/transl_except=(pos:<base_range>,aa:<amino_acid>)
/usedin=accnum:feature_label

Old GB key: sigp; pept

Old EMBL key: CDS

Comment: /codon_start has valid value of the single base location of a codon start;
/codon is used to specify unusual genetic codes, including rare-usage start codons, organellar codes, etc.;
/codon is used to describe a single codon exception to the code defined as the "normal" code for the organism; it implies that the translation it specifies is used throughout the feature
/codon specifies that a specific codon specified by "seq" codes for the amino acid or stop codon specified by "aa";
/transl_except is used to specify a single codon the translation of which does not conform to the genetic code defined by the "normal" code for organism or by the /codon qualifiers given for the feature
can this key be applied to non-organellar translocated proteins?

Feature Key: transposon

Definition: transposable element. TN; a DNA sequence able to replicate and insert one copy at (or, without replication, to move itself to) a new location in the genome

Optional qualifiers:
/citation=[number]
/evidence=<evidence_value>
/function="text"
/label=feature_label
/note="text"
/organism="text"
/partial
/standard_name="text"
/usedin=accnum:feature_label

Molecule scope: DNA

Old GB key: trns

Old EMBL key: TPOSON

Comment: the transposon name should be given using the /standard_name qualifier; can the /organism qualifier be used to name the natural host organism?

Feature Key	tRNA
Definition	mature transfer RNA, a small RNA molecule (75-85 bases long) that mediates the translation of a nucleic acid sequence into an amino acid sequence
Optional qualifiers	/anticodon=(pos:<base_range>,aa:<amino_acid>) /citation=[number] /evidence=<evidence_value> /function="text" /gene="text" /label=feature_label /note="text" /partial /product="text" /pseudo /standard_name="text" /usedin=accnum:feature_label
Old GB key	tRNA
Old EMBL key	TRNA

Feature Key	unsure
Definition	author is unsure of exact sequence in this region
Optional qualifiers	/citation=[number] /usedin=accnum:feature_label /label=feature_label /note="text"
Old GB key	unsure
Old EMBL key	UNSURE
Comment	an alternative sequence at this site can be specified in the location as follows: replace(seq_location,change_loc) where change_loc is an alternative sequence

Feature Key	variation
Definition	a related strain contains stable mutations from the same gene (e.g., RFLPs, polymorphisms, etc.) which differ from the presented sequence at this location (and possibly others).
Optional qualifiers	/citation=[number] /frequency="text" /gene="text" /label=feature_label /note="text" /phenotype="text"

/product="text"
/standard_name="text"
/type="text"
/usedin=accnum:feature_label

Old GB key variant

Old EMBL key VARIANT

Comment format of location is: replace(seq_location,variation) where seq_location is a location descriptor pointing to a region of the presented sequence and variation is a location descriptor (typically a literal sequence) with which the sequence at seq_location is replaced in the variant. The /type qualifier may be used to specify a strain name

Feature Key **virion**

Definition viral genomic sequence as it is encapsidated, as distinguished from its proviral form (integrated in a host cell's chromosome)

Optional qualifiers /citation=[number]
/label=feature_label
/note="text"
/organism="text"
/partial
/usedin=accnum:feature_label

Organism scope Viruses

Old GB key virion

Feature Key **3'clip**

Definition 3'-most region of a precursor transcript that is clipped off during processing

Optional qualifiers /citation=[number]
/evidence=<evidence_value>
/function="text"
/label=feature_label
/note="text"
/partial
/standard_name="text"
/usedin=accnum:feature_label

Old GB key mRNA

Old EMBL key MSG

Feature Key 3'UTR

Definition region near or at the 3' end of a mature transcript (usually following the stop codon) that is not translated into a protein; trailer

Optional qualifiers
/citation=[number]
/evidence=<evidence_value>
/function="text"
/label=feature_label
/note="text"
/partial
/standard_name="text"
/usedin=accnum:feature_label

Old GB key mRNA

Old EMBL key MSG

Feature Key 5'clip

Definition 5'-most region of a precursor transcript that is clipped off during processing

Optional qualifiers
/citation=[number]
/function="text"
/evidence=<evidence_value>
/label=feature_label
/note="text"
/partial
/standard_name="text"
/usedin=accnum:feature_label

Old GB key mRNA

Old EMBL key MSG

Feature Key 5'UTR

Definition region near or at the 5' end of a mature transcript (usually preceding the initiation codon) that is not translated into a protein; leader

Optional qualifiers
/citation=[number]
/evidence=<evidence_value>
/function="text"
/label=feature_label
/note="text"
/partial
/standard_name="text"
/usedin=accnum:feature_label

Old GB key mRNA

Old EMBL key	MSG

Feature Key **-10_signal**

Definition	Pribnow box; a conserved region about 10 bp upstream of the start point of bacterial transcription units which may be involved in binding RNA polymerase; consensus=TAtAaT [1,2,3,4]
Optional qualifiers	/citation=[number] /label=feature_label /note="text" /standard_name="text" /usedin=accnum:feature_label
Organism scope	prokaryotes
Molecule scope	DNA
Old EMBL key	PRM
References	[1] Schaller, H., Gray, C., and Hermann, K. Proc Natl Acad Sci USA **72**, 737-741 (1974) [2] Pribnow, D. Proc Natl Acad Sci USA **72**, 784-788 (1974) [3] Hawley, D.K. and McClure, W.R. "Compilation and analysis of Escherichia coli promoter DNA sequences" Nucl Acid Res **11**, 2237-2255 (1983) [4] Rosenberg, M. and Court, D. "Regulatory sequences involved in the promotion and termination of RNA transcription" Ann Rev Genet **13**, 319-353 (1979)

Feature Key **-35_signal**

Definition	a conserved hexamer about 35 bp upstream of the start point of bacterial transcription units; consensus=TTGACa [ ] or TGTTGACA [ ]
Optional qualifiers	/citation=[number] /label=feature_label /note="text" /usedin=accnum:feature_label
Organism scope	prokaryotes
Molecule scope	DNA
Old EMBL key	PRM
References	[1] Takanami, M., et al. Nature **260**, 297-302 (1976) [2] Moran, C.P., Jr., et al. Molec Gen Genet **186**, 339-346 (1982) [3] Maniatis, T., et al. Cell **5**, 109-113 (1975)
Comment	note: original signal sequence proposed was TGTTG [4]

7.4 Appendix IV Summary of qualifiers for feature keys

The following is a list of available qualifiers for feature keys and their usage. The information is arranged as follows:

Qualifier	name of qualifier; qualifier requires a value if followed by an equal sign
Definition	definition of the qualifier
Value format	format of value, if required
Example	example of qualifier with value
Comment	comments, questions and clarifications

Qualifier	**/anticodon=(pos: ,aa:)**
Definition	location of the anticodon of tRNA and the amino acid for which it codes
Value format	pos:<base_range>,aa:<amino_acid> where base_range is the position of the anticodon and amino_acid is the abbreviation for the amino acid encoded
Example	/anticodon=(pos:34..36, aa:Phe)

Qualifier	**/bound_moiety=**
Definition	moiety bound
Value format	"text"
Example	/bound_moiety="repressor"

Qualifier	**/citation=**
Definition	reference to a citation listed in the entry reference field
Value format	[integer-number] where integer-number is the number of the reference as enumerated in the reference field
Example	/citation=[3]
Comment	used to indicate the citation providing the claim of and/or evidence for a feature; brackets are used for conformity

Qualifier	**/codon=(seq: ,aa:)**
Definition	specifies a codon which is different from any found in the reference genetic code
Value format	(seq:"codon-sequence",aa:<amino_acid>) where codon-sequence is the bases of the codon and amino_acid is the abbreviation for the translated amino acid
Example	/codon=(seq:"ttt", aa:Leu)

Comment	Amino acids not on the controlled vocabulary list can be indicated by using "aa:OTHER" as the amino acid designation and giving the name of the residue in a /note

Qualifier	**/codon_start=**
Definition	protein coding region reading frame relative to first base number in location
Value format	position
Example	/codon_start=213

Qualifier	**/cons_splice=**
Definition	differentiates between intron splice sites that conform to the 5'-GT ... AG-3' splice site consensus
Value format	(5'site:<bool>, 3'site:<bool>)
Example	/cons_splice=(5'site:YES, 3'site:NO)
Comment	valid boolean values are **yes** and **no**; since the vast majority of splice sites conform to the consensus, this qualifier should be used only when one does not and the sequence has been checked

Qualifier	**/direction=**
Definition	direction of DNA replication
Value format	**left**, **right**, or **both** where **left** indicates toward the 5' end of the entry sequence (as presented) and **right** indicates toward the 3' end
Example	/direction=LEFT

Qualifier	**/EC_number=**
Definition	Enzyme Commission number for enzyme product of sequence
Value format	"text"
Example	/EC_number=1.1.2.4
Comment	Valid values for EC numbers are defined in the list prepared by the IUPAC-IUB Commission on Biochemical Enzyme Nomenclature (published in *Enzyme Nomenclature 1984*. New York: Academic Press (1984) or a more recent revision thereof)

Qualifier	**/evidence=**
Definition	value indicating the nature of supporting evidence
Value format	**experimental**
Example	/evidence=EXPERIMENTAL
Comment	**experimental** is the only currently valid value for the /evidence qualifier. **experimental** indicates that the feature identification or assignment is supported by direct experimental (rather than or in addition to pattern

similarity) evidence. see Feature Table Annotation Standards Guide for application guidelines and copious examples

Qualifier	**/frequency=**
Definition	frequency of the occurrence of a feature
Value format	text representing the fraction of population carrying the variation expressed as a decimal fraction
Example	/frequency=0.85

Qualifier	**/function=**
Definition	function attributed to a sequence
Value format	"text"
Example	function="essential for recognition of cofactor"
Comment	/function is used when the gene name and/or product name do not convey the function attributable to a sequence

Qualifier	**/gene=**
Definition	symbol of the gene corresponding to a sequence region
Value format	"text"
Example	/gene="ilvE"
Comment	see O'Brien, S.J., ed., *Genetic Maps 1987*, Cold Spring Harbor or a recent revision. Are there other standard genetic maps or symbol lists which should be included?

Qualifier	**/label=**
Definition	a label used to permanently tag a feature
Value format	feature_label. feature labels follow the naming conventtions for all feature table objects (see Sections 3.1 and 3.4)
Example	/label=Alb1_exon1

Qualifier	**/mod_base=**
Definition	abbreviation for a modified nucleotide base
Value format	modified_base
Example	/mod_base=m5c
Comment	Modified nucleotides not found in the restricted vocabulary list can be annotated by entering '/mod_base=OTHER' with '/note="name of modified base"'

Qualifier	**/note=**
Definition	any comment or additional information
Value format	"text"
Example	/note="This qualifier is equivalent to a comment."

Qualifier	**/number=**
Definition	a number to indicate the order of genetic elements (e.g., exons or introns) in the 5' to 3' direction
Value format	<integer>
Example	/number=4

Qualifier	**/organism=**
Definition	name of organism if different from that contained in the entry ORGANISM (OS) field
Value format	"text"
Example	/organism="Homo sapiens"
Comment	Organism names are controlled vocabulary. Quotation marks are used because they may contain a variety of symbols which may confuse a parser.

Qualifier	**/partial**
Definition	differentiates between complete regions and partial ones
Value format	none
Example	/partial
Comment	need be used only with features where incompleteness is not indicated by the location given (i.e., the location does not contain a '<' or '>')

Qualifier	**/phenotype=**
Definition	phenotype conferred by the feature
Value format	"text"
Example	/phenotype="erythromycin resistance"

Qualifier	**/product=**
Definition	name of a product encoded by a sequence
Value format	"text"
Example	/product="catalase"
Comment	

Qualifier	/pseudo
Definition	indicates that this feature is a non-functional version of the element named by the feature key
Value format	none
Example	/pseudo

Qualifier	/rpt_family=
Definition	type of repeated sequence; "Alu" or "Kpn", for example
Value format	"text"
Example	/rpt_family="Alu"
Comment	preferred usage is to qualify the repeat_region instead of any of the constituent repeat_units

Qualifier	/rpt_type=
Definition	organization of repeated sequence
Value format	**tandem, inverted, flanking, terminal, direct, dispersed,** and **other**
Example	/rpt_type=INVERTED
Comment	preferred usage is to qualify the repeat_region instead of any of the constituent repeat_units. definitions of these values will be added in a future release of this document. see Singer, M. Int Rev Cytol **76**, 67-112 (1982); Cell **26**, 293-95 (1981); Hardman, N. Biochem J **234**, 1-11 (1986)

Qualifier	/rpt_unit=
Definition	identity of repeat unit which constitutes a repeat_region
Value format	<feature_label> *or* <base_range>
Example	/rpt_unit=Alu_rpt1 /rpt_unit=202..245
Comment	used to indicate feature which defines (or base range of) the repeat unit of which a repeat region is made

Qualifier	/standard_name=
Definition	accepted standard name for this feature
Value format	"text"
Example	/standard_name="dotted"
Comment	use /standard_name to give full gene name, but use /gene to give gene symbol (in the above example /gene="Dt")

Qualifier	**/transl_except=(pos: ,aa:)**
Definition	translational exception: single codon the translation of which does not conform to genetic code defined by Organism and /codon=
Value format	(pos:base_range,aa:<amino_acid>) where amino_acid is the amino acid coded by the codon at the base_range position
Example	/transl_except=(pos:213..216, aa:Trp)
Comment	If the amino acid is not on the restricted vocabulary list use, e.g., '/transl_except=(pos:213..215, aa:OTHER)' with '/note="name of unusual amino acid" '.

Qualifier	**/type=**
Definition	name of a strain if different from that contained in the entry SOURCE field
Value format	"text"
Example	/type="W64msw"

Qualifier	**/usedin=**
Definition	indicates that the feature is used in a compound feature in another entry
Value format	Accession-number:feature-name *or* Database_name::Acc_number:feature_label
Example	/usedin=X10087:proteinx
Comment	Database_name is an abbreviation for the name of the database in which the entry for the Acc-number accession number can be found.

7.5 Appendix V Controlled vocabularies

This appendix contains information on the restricted vocabulary fields used in the Feature Table. The information contained in this appendix is subject to change, please contact the database staff for the most recent information concerning controlled vocabularies. This appendix is organized as follows:

Authority	The organization with authority to define the vocabulary
Reference	Publications of (or about) the vocabulary
Contact	Name and location of database staff member responsible for maintaining the database copy of the vocabulary
Scope	Feature Table qualifiers which take members of this vocabulary as values
Listing	A listing of the current vocabulary with definitions or explanations

This appendix includes reference lists for the following controlled vocabulary fields:

Nucleotide base codes (IUPAC)

Modified base abbreviations

Amino acid abbreviations (including modified and unusual amino acids)

7.5.1 Nucleotide base codes (IUPAC)

Authority	Nomenclature Committee of the International Union of Biochemistry
Reference	Cornish-Bowden, A. Nucl Acid Res **13**, 3021-3030 (1985)
Contact	David Hazledine, EMBL
Scope	Location descriptors, /codon

Listing

Symbol	Meaning
a	a; adenine
c	c; cytosine
g	g; guanine
t	t; thymine in DNA; uracil in RNA
m	a or c
r	a or g
w	a or t
s	c or g
y	c or t
k	g or t
v	a or c or g; not t
h	a or c or t; not g
d	a or g or t; not c
b	c or g or t; not a
n	a or c or g or t

7.5.2 Modified base abbrerviations

Authority	Sprinzl, M. and Gauss, D.H.
Reference	Sprinzl, M. and Gauss, D.H. Nucl Acid Res **10**, r1 (1982). (note that in Cornish_Bowden, A. Nucl Acid Res **13**, 3021-3030 (1985) the IUPAC-IUB declined to recommend a set of abbreviations for modified nucleotides)
Contact	Jamie Hayden, LANL
Scope	/mod_base

Listing

Abbre-viation	Modified base description
ac4c	4-acetylcytidine
chm5u	5-(carboxyhydroxylmethyl)uridine
cm	2'-O-methylcytidine
cmnm5s2u	5-carboxymethylaminomethyl-2-thiouridine
cmnm5u	5-carboxymethylaminomethyluridine
d	dihydrouridine
fm	2'-O-methylpseudouridine
gal q	beta,D-galactosylqueosine
gm	2'-O-methylguanosine
i	inosine
i6a	N6-isopentenyladenosine
m1a	1-methyladenosine
m1f	1-methylpseudouridine
m1g	1-methylguanosine
m1i	1-methylinosine
m22g	2,2-dimethylguanosine
m2a	2-methyladenosine
m2g	2-methylguanosine
m3c	3-methylcytidine
m5c	5-methylcytidine
m6a	N6-methyladenosine
m7g	7-methylguanosine
mam5u	5-methylaminomethyluridine
mam5s2u	5-methoxyaminomethyl-2-thiouridine
man q	beta,D-mannosylqueosine
mcm5s2u	5-methoxycarbonylmethyl-2-thiouridine
mcm5u	5-methoxycarbonylmethyluridine
mo5u	5-methoxyuridine
ms2i6a	2-methylthio-N6-isopentenyladenosine
ms2t6a	N-((9-beta-D-ribofuranosyl-2-methyltiopurine-6-yl)carbamoyl)threonine
mt6a	N-((9-beta-D-ribofuranosylpurine-6-yl)N-methyl-carbamoyl)threonine
mv	uridine-5-oxyacetic acid-methylester
o5u	uridine-5-oxyacetic acid (v)
osyw	wybutoxosine
p	pseudouridine
q	queosine
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
yw	wybutosine
x	3-(3-amino-3-carboxypropyl)uridine, (acp3)u
OTHER	*(requires /note= qualifier)*

7.5.3 Amino acid abbreviations

Authority	IUPAC-IUB Commission on Biological Nomenclature.
Reference	IUPAC-IUB Commission on Biological Nomenclature. J. Biol Chem **241**, 2491 (1966); IUPAC-IUB Commission on Biological Nomenclature. J. Biol Chem **243**, 3557 (1968)
Contact	Patricia Kahn, EMBL
Scope	/anticodon, /codon, /transl_except

Listing (note that the abbreviations are legal values for amino acids, not the full names)

Abbreviation	Amino acid name
Ala	Alanine
Arg	Arginine
Asn	Asparagine
Asp	Aspartic acid (Aspartate)
Cys	Cysteine
Gln	Glutamine
Glu	Glutamic acid (Glutamate)
Gly	Glycine
His	Histidine
Ile	Isoleucine
Leu	Leucine
Lys	Lysine
Met	Methionine
Phe	Phenylalanine
Pro	Proline
Ser	Serine
Thr	Threonine
Trp	Tryptophan
Tyr	Tyrosine
Val	Valine
TERM	termination codon

Modified and unusual Amino Acids

Abbreviation	Amino acid
Aad	2-Aminoadipic acid
bAad	3-Aminoadipic acid
bAla	beta-Alanine, beta-Aminoproprionic acid
Abu	2-Aminobutyric acid
4Abu	4-Aminobutyric acid, piperidinic acid
Acp	6-Aminocaproic acid
Ahe	2-Aminoheptanoic acid
Aib	2-Aminoisobutyric acid
bAib	3-Aminoisobutyric acid
Apm	2-Aminopimelic acid
Dbu	2,4-Diaminobutyric acid
Des	Desmosine
Dpm	2,2'-Diaminopimelic acid
Dpr	2,3-Diaminoproprionic acid
EtGly	N-Ethylglycine
EtAsn	N-Ethylasparagine
Hyl	Hydroxylysine
aHyl	allo-Hydroxylysine
3Hyp	3-Hydroxyproline
4Hyp	4-Hydroxyproline
Ide	Isodesmosine
aIle	allo-Isoleucine
MeGly	N-Methylglycine, sarcosine
MeIle	N-Methylisoleucine
MeLys	6-N-Methyllysine
MeVal	N-Methylvaline
Nva	Norvaline
Nle	Norleucine
Orn	Ornithine
OTHER	*(requires /note=)*