STANDARD ST.26

SEPTEMBER 2017 CHANGES

MAIN BODY
   Editorial Note, revised June 2017 - CWS/5
   INTRODUCTION, SCOPE, REFERENCES, REPRESENTATION OF SEQUENCES and STRUCTURE OF THE SEQUENCE LISTING IN XML, revised June 2017 - CWS/5

ANNEX I
   Sections 2, 3, 5, 6, 7, 8 and 9, revised June 2017 - CWS/5

ANNEX II
   Revised June 2017 - CWS/5

ANNEX III
   Revised June 2017 - CWS/5

ANNEX VI
   Added June 2017 - CWS/5
STANDARD ST.26

RECOMMENDED STANDARD FOR THE PRESENTATION OF NUCLEOTIDE AND AMINO ACID SEQUENCE LISTINGS USING XML (EXTENSIBLE MARKUP LANGUAGE)

Version 1.01.1

Adopted Approved by the Committee on WIPO Standards (CWS) at its reconvened fourth session on March 24, 2016, fifth session on June 2, 2017.

Editorial Note prepared by the International Bureau

At its fifth session, the Committee on WIPO Standards (CWS) agreed that the transition from WIPO Standard ST.25 to Standard ST.26 takes place in January 2022. Meanwhile, Standard ST.25 should continue to be used.

The Committee on WIPO Standards (CWS) agreed to ask industrial property offices to postpone the preparations for implementation of this new WIPO Standard ST.26 until the recommendations for the transition from WIPO Standard ST.25 to the new Standard ST.26 is agreed on by the CWS at its next session to be held in 2017. Meanwhile, Standard ST.25 should continue to be used.

The Standard is published for information purposes of industrial property offices and other interested parties.
TABLE OF CONTENTS

INTRODUCTION ................................................................................................................. 4
DEFINITIONS ....................................................................................................................... 4
SCOPE ................................................................................................................................ 5
REFERENCES ..................................................................................................................... 6
PRESENTATION OF SEQUENCES ....................................................................................... 6
  Nucleotide sequences ...................................................................................................... 6
  Amino acid sequences .................................................................................................... 9
  Presentation of special situations .................................................................................. 11
STRUCTURE OF THE SEQUENCE LISTING IN XML .......................................................... 11
  Root element .................................................................................................................. 12
  General information part ............................................................................................... 13
  Sequence data part ......................................................................................................... 16
  Feature table ................................................................................................................ 18
  Feature keys .................................................................................................................. 18
  Mandatory feature keys ............................................................................................... 18
  Feature location ............................................................................................................ 18
  Feature qualifiers ......................................................................................................... 20
  Mandatory feature qualifiers ..................................................................................... 20
  Qualifier elements ....................................................................................................... 20
  Free text ....................................................................................................................... 22
  Coding sequences ....................................................................................................... 22
  Variants ....................................................................................................................... 23

ANNEXES
Annex I - Controlled vocabulary
Annex II - Document Type Definition for Sequence Listing (DTD)
Annex III - Sequence Listing Specimen (XML file)
Annex IV - Character Subset from the Unicode Basic Latin Code Table
Annex V - Additional data exchange requirements (for patent offices only)
Annex VI - Guidance document
  Appendix - Guidance document sequences in XML
INTRODUCTION

1. This Standard defines the nucleotide and amino acid sequence disclosures in a patent application required to be included in a sequence listing, the manner in which those disclosures are to be characterized, represented, and the Document Type Definition (DTD) for a sequence listing in XML (eXtensible Markup Language). It is recommended that industrial property offices accept any sequence listing compliant with this Standard filed as part of a patent application or in relation to a patent application.

2. The purpose of this Standard is to:
   (a) allow applicants to draw up a single sequence listing in a patent application acceptable for the purposes of both international and national or regional procedures;
   (b) enhance the accuracy and quality of presentations of sequences for easier dissemination, benefiting applicants, the public and examiners;
   (c) facilitate searching of the sequence data; and
   (d) allow sequence data to be exchanged in electronic form and introduced into computerized databases.

DEFINITIONS

3. For the purpose of this Standard, the expression:
   (a) "amino acid" means any amino acid that can be represented using any of the symbols set forth in Annex I (see Section 3, Table 3). Such amino acids include, inter alia, D-amino acids and amino acids containing modified or synthetic side chains. Amino acids will be construed as unmodified L-amino acids unless further described in the feature table as modified according to paragraph 30. For the purpose of this standard, a peptide nucleic acid (PNA) residue is not considered an amino acid, but is considered a nucleotide as set forth in paragraph 3(g)(i)(2).
   (b) "controlled vocabulary" is the terminology contained in this Standard that must be used when describing the features of a sequence, i.e., annotations of regions or sites of interest as set forth in Annex I.
   (c) "enumeration of its residues" means disclosure of a sequence in a patent application by listing, in order, each residue of the sequence, wherein:
      (i) the residue is represented by a name, abbreviation, symbol, or structure (e.g., HHHHHHQ or HisHisHisHisHisGln); or
      (ii) multiple residues are represented by a shorthand formula (e.g., HisGln).
   (d) "intentionally skipped sequence", also known as an empty sequence, refers to a placeholder to preserve the numbering of sequences in the sequence listing for consistency with the application disclosure, for example, where a sequence is deleted from the disclosure to avoid renumbering of the sequences in both the disclosure and the sequence listing.
   (e) "modified amino acid" means any amino acid as described in paragraph 3[a] other than L-alanine, L-arginine, L-asparagine, L-aspartic acid, L-cysteine, L-glutamine, L-glutamic acid, L-histidine, L-isoleucine, L-leucine, L-lysine, L-methionine, L-phenylalanine, L-proline, L-pyrrolidine, L-serine, L-selenocysteine, L-threonine, L-tryptophan, L-tyrosine, or L-valine.
4. For the purpose of this Standard, the word(s):

(a) "may" refers to an optional or permissible approach, but not a requirement.
(b) "must" refers to a requirement of the Standard; disregard of the requirement will result in noncompliance.
(c) "must not" refers to a prohibition of the Standard.
(d) "should" refers to a strongly encouraged approach, but not a requirement.
(e) "should not" refers to a strongly discouraged approach, but not a prohibition.

SCOPE

5. This Standard establishes the requirements for the presentation of nucleotide and amino acid sequence listings of sequences disclosed in patent applications.

---

Ref.: Standards - ST.26

en / 03-26-01

Date: June 2016

September 2017
6. A sequence listing complying with this Standard (hereinafter sequence listing) contains a general information part and a sequence data part. The sequence listing must be presented as a single file in XML using the Document Type Definition (DTD) presented in Annex II. The purpose of the bibliographic information contained in the general information part is solely for association of the sequence listing to the patent application for which the sequence listing is submitted. The sequence data part is composed of one or more sequence data elements each of which contain information about one sequence. The sequence data elements include various feature keys and subsequent qualifiers based on the International Nucleotide Sequence Database Collaboration (INSDC) and UniProt specifications.

7. For the purpose of this Standard, a sequence for which inclusion in a sequence listing is required is one that is disclosed anywhere in an application by enumeration of its residues and can be represented as:
   (a) an unbranched sequence or a linear portion of a branched sequence containing ten or more specifically defined nucleotides, wherein adjacent nucleotides are joined 3’ to 5’ (or 5’ to 3’), or
   (i) a 3’ to 5’ (or 5’ to 3’) phosphodiester linkage; or
   (ii) any chemical bond that results in an arrangement of adjacent nucleobases that mimics the arrangement of nucleobases in naturally occurring nucleic acids; or
   (b) an unbranched sequence or a linear portion of a branched sequence containing four or more specifically defined amino acids, wherein adjacent amino acids are joined by peptide bonds.

8. A sequence listing must not include, as a sequence assigned its own sequence identification number, any sequences having fewer than ten specifically defined nucleotides, or fewer than four specifically defined amino acids.

REFERENCES

9. References to the following Standards and resources are of relevance to this Standard:

   - International Nucleotide Sequence Database Collaboration (INSDC) http://www.insdc.org/;
   - UniProt Consortium http://www.uniprot.org/;
   - W3C XML 1.0 http://www.w3.org/;
   - WIPO Standard ST.2 Standard Manner for Designating Calendar Dates by Using the Gregorian Calendar;
   - WIPO Standard ST.3 Two-Letter Codes for the Representation of States, Other Entities and Intergovernmental Organizations;
   - WIPO Standard ST.16 Identification of different kinds of patent documents;
   - WIPO Standard ST.25 Presentation of nucleotide and amino acid sequence listings.

PRESENTATION OF SEQUENCES

10. Each sequence encompassed by paragraph 7 must be assigned a separate sequence identification number, including a sequence which is identical to a region of a longer sequence. The sequence identification numbers must begin with number 1, and increase consecutively by integers. Where no sequence is present for a sequence identification number, i.e. an intentionally skipped sequence, “000” must be used in place of a sequence (see paragraph 58). The total number of sequences must be indicated in the sequence listing and must equal the total number of sequence identification numbers, whether followed by a sequence or by “000.”

Nucleotide sequences

11. A nucleotide sequence must be presented only by a single strand, in the 5’-end to 3’-end direction from left to right, or in the direction from left to right that mimics the 5’-end to 3’-end direction. The designations 5’ and 3’ or any other similar designations must not be included in the sequence. A double-stranded nucleotide sequence disclosed by enumeration of the residues of both strands must be presented as:
   (a) a single sequence or as two separate sequences, each assigned its own sequence identification number, where the two separate strands are fully complementary to each other, or
(b) two separate sequences, each assigned its own sequence identification number, where the two strands are not fully complementary to each other.

12. For the purpose of this Standard, numbering of positions must start at the first base of the sequence if the sequence is linear. It must be continuous through the whole sequence in the direction 5' to 3'. When the sequence is circular in configuration, the above numbering method will be used. If the sequence is circular in configuration, the directionality must be specified, indicating whether the clockwise or counterclockwise direction begins with the first base. When nucleotide sequences are also applicable to nucleotide sequences that are circular in configuration, the directionality must be specified, indicating whether the clockwise or counterclockwise direction begins with the first base.

13. All nucleotides in a sequence must be represented using the symbols set forth in Annex I (see Section 1, Table 1). Only lower case letters must be used. Any symbol used to represent a nucleotide is the equivalent of only one residue.

14. The symbol “t” will be construed as thymine in DNA and uracil in RNA. Uracil in DNA or thymine in RNA is considered a modified nucleotide and must be accompanied by a further description as provided in paragraph 16.
Example 2: Modified nucleotide “xanthine” using “OTHER” from Annex I (see Section 2, Table 2)

Example 3: A nucleotide sequence composed of modified nucleotides encompassed by paragraph 3(g)(ii)(2) with two individual nucleotides that include a further modification

21. Any “unknown” nucleotide must be represented by the symbol “n” in the sequence. An “unknown” nucleotide should be further described in the feature table (see paragraph 60 et seq.) using the feature key “unsure”. The symbol “n” is the equivalent of only one residue.
22. A region containing a known number of contiguous "a", "c", "g", "t", or "n" residues for which the same description applies may be jointly described using a single INSDFeature element with the syntax "x..y" as the location descriptor in the element INSDFeature_location. For the presentation of sequence variants, i.e., deletions, insertions or substitutions, see paragraphs 92 to 98.

23. The following example illustrates the presentation of a region of modified nucleotides for which the same description applies, according to paragraph 22 above:

```
<INSDFeature>
  <INSDFeature_key>modified_base</INSDFeature_key>
  <INSDFeature_location>358..485</INSDFeature_location>
  <INSDFeature_quals>
    <INSDFeature_key>modified_base</INSDFeature_key>
    <INSDFeature_location>358..485</INSDFeature_location>
  </INSDFeature_quals>
</INSDFeature>
```

Amino acid sequences

24. The amino acids in a protein or peptide amino acid sequence must be represented in the amino to carboxy direction from left to right. The amino and carboxy groups must not be represented in the sequence.

25. For the purpose of Numbering amino acid positions, the first amino acid of the sequence, with residue position number 1, including amino acids preceding the mature protein, for example, pre-sequences, pro-sequences, pre-pro-sequences and signal sequences, must be contiguous through the whole sequence. Numbering is continuous through the entire sequence in the amino to carboxy direction.

26. All amino acids in a sequence must be represented using the symbols set forth in Annex I (see Section 3, Table 3). Only upper case letters must be used. Any symbol used to represent an amino acid is the equivalent of only one residue.

27. Where an ambiguity symbol (representing two or more amino acids in the alternative) is appropriate, the most restrictive symbol should be used. For example, if an amino acid in a given position could be aspartic acid or asparagine, the symbol "B" should be used, rather than "X". The symbol "X" will be construed as any one of "A", "R", "N", "D", "C", "Q", "E", "G", "H", "I", "L", "K", "M", "F", "P", "Q", "O", "S", "U", "T", "W", "Y", or "V", except where it is used with a further description in the feature table, as provided by paragraphs 92 to 98.

28. Disclosed amino acid sequences separated by one or more blank spaces or internal terminator symbols, represented for example by "Ter" or asterisk ** or period .. in a disclosure or a blank space, must be presented as separate sequences for each amino acid sequence that contains at least four specifically defined amino acids and is encompassed by paragraph 47. Each such separate sequence must be presented in the sequence listing with assigned its own sequence identification number, including the symbols set forth in Annex I (see Sections 3, Table 3). Terminus symbols and spaces must not be included in sequences in a sequence listing (see paragraph 57).

29. Modified amino acids, including D-amino acids, should be represented in the sequence as the corresponding unmodified amino acids whenever possible. Any modified amino acid in a sequence that cannot otherwise be represented by any symbol in Annex I (see Section 3, Table 3), i.e., an "other amino acid", must be represented by "X". The symbol "X" is the equivalent of only one residue.

30. A modified amino acid must be further described in the feature table (see paragraph 60 et seq.). Where applicable, the feature keys "CARBOHYD" or "LIPID" should be used together with the qualifier "NOTE". The feature key "MOD_RES" should be used for post-translationally modified amino acids together with the qualifier "NOTE"; otherwise the feature key "SITE" together with the qualifier "NOTE" should be used. The value for the qualifier "NOTE" must either be an abbreviation set forth in Annex I (see Section 4, Table 4), or the complete, unabbreviated name of the modified amino acid. The abbreviations set forth in Table 4 referred to above or the complete, unabbreviated names must not be used in the sequence itself.

31. The following examples illustrate the presentation of modified amino acids according to paragraph 22 above:
Example 1: Post-translationally modified amino acid

```xml
<INSDFeature>
    <INSDFeature_key>MOD_RES</INSDFeature_key>
    <INSDFeature_location>3</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>Hyp</INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
</INSDFeature>
```

Example 2: Non post-translationally modified amino acid

```xml
<INSDFeature>
    <INSDFeature_key>SITE</INSDFeature_key>
    <INSDFeature_location>3</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>Orn</INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
</INSDFeature>
```

Example 3: D-amino acid

```xml
<INSDFeature>
    <INSDFeature_key>SITE</INSDFeature_key>
    <INSDFeature_location>9</INSDFeature_location>
    <INSDQualifier>
        <INSDQualifier_name>NOTE</INSDQualifier_name>
        <INSDQualifier_value>D-Arginine</INSDQualifier_value>
    </INSDQualifier>
</INSDFeature>
```

32. Any "unknown" or "other" amino acid not covered by paragraph 28 must be represented by the symbol "X" in the sequence. The symbol "X" is the equivalent of only one residue. An "unknown" amino acid designated as "X" must be further described in the feature table (see paragraph 60 et seq.) using the feature key "UNSURE" and optionally the qualifier "NOTE". An "other" amino acid designated as "X" must be further described using the feature key "SITE" or "MOD_RES" as appropriate, and the qualifier "NOTE" with the complete, unabbreviated name of the "other" amino acid. The symbol "X" is the equivalent of only one residue.

33. The following examples illustrate the presentation of an "unknown" or "other" amino acid according to paragraphs 31 and paragraph 32 above:

Example 1: "unknown" amino acid

```xml
<INSDFeature>
    <INSDFeature_key>UNSURE</INSDFeature_key>
    <INSDFeature_location>3</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>A or V</INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
</INSDFeature>
```

Example 2: "other" amino acid
34. A region containing a known number of contiguous “X” residues for which the same description applies may be jointly described using the syntax “x..y” as the location descriptor in the element INSDFeature_location (see paragraphs 35.4 to 35.6). For representation of sequence variants, i.e., deletions, insertions, or substitutions, see paragraphs 35.9 to 35.12.

Presentation of special situations

35. A sequence disclosed by enumeration of its residues that is constructed as a single continuous sequence from one or more non-contiguous segments of a larger sequence or of segments from different sequences must be included in the sequence listing as a single sequence with a single and assigned its own sequence identification number.

36. A sequence disclosed by enumeration of its residues that contains regions of specifically enumerated residues separated by one or more regions of contiguous “n” or “X” residues (see paragraphs 15 and 27, respectively), wherein the exact number of “n” or “X” residues in each region is disclosed, must be included in the sequence listing as a single sequence with a single and assigned its own sequence identification number.

37. A sequence disclosed by enumeration of its residues that contains regions of specifically enumerated residues separated by one or more gaps of an unknown or undisclosed number of residues must not be represented in the sequence listing as a single sequence. Each region of specifically defined residues that is encompassed by paragraph 7 must be included in the sequence listing as multiple separate sequences. Each such separate sequence must contain one region of specifically enumerated residues with sequence and assigned its own sequence identification number. Where the number of separate sequences is equal to the number of regions of specifically enumerated residues. Sequences containing gaps of an unknown or undisclosed number of residues must not be included in the sequence listing as a single sequence.

STRUCTURE OF THE SEQUENCE LISTING IN XML

38. In accordance with paragraph 33 above, an XML instance of a sequence listing file according to this Standard is composed of:

(a) general information part, which contains information concerning the patent application to which the sequence listing is directed; and
(b) sequence data part, which contains one or more sequence data elements, each of which, in turn contain information about one sequence.

An example of a sequence listing is provided in Annex III.

39. The sequence listing must be presented in XML 1.0 using the DTD presented in the Annex II “Document Type Definition for Sequence Listing”.

(a) The first line of the XML instance must contain the XML declaration:

```xml
<?xml version="1.0" encoding="UTF-8"?>
```

(b) The second line of the XML instance must contain a document type (DOCTYPE) declaration:

```xml
<!DOCTYPE ST26SequenceListing PUBLIC "//WIPO/DTD Sequence Listing 1.0//EN"
"ST26SequenceListing_V1_0.dtd">
```

40. The entire electronic sequence listing must be contained within one file. The file must be encoded using Unicode UTF-8, with the following restrictions:

(a) the information contained in the elements ApplicantName, InventorName and InventionTitle of the general information part, may be composed of any Unicode characters except the reserved characters, which must be replaced as set forth in paragraph 41; and
(b) the information contained in all other elements of the general information part and in all elements of the sequence data part must be composed of printable characters (including the space character) from the Unicode Basic Latin code table excluding the reserved characters, which must be replaced as set forth in paragraph 41, (i.e., limited to Unicode code points 0020, 0021, 0023 through 0026, 0028 through 003B, 003D, and 003F through 007E – see Annex IV), and the only character entities permitted are the predefined entities set forth in paragraph 41.

41. In an XML instance of a sequence listing, the following reserved characters must be replaced by the corresponding predefined entities when used in a value of an attribute or content of an element:

<table>
<thead>
<tr>
<th>Reserved Character</th>
<th>Predefined Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>&gt;</td>
<td>&gt;</td>
</tr>
<tr>
<td>&amp;</td>
<td>&amp;</td>
</tr>
<tr>
<td>'</td>
<td>&quot;</td>
</tr>
<tr>
<td>'</td>
<td>'</td>
</tr>
</tbody>
</table>

See paragraph 4271 for an example.

42. All mandatory elements must be populated (except as provided for in paragraph 58 for an intentionally skipped sequence). Optional elements for which content is not available should not appear in the XML instance (except as provided for in paragraph 95 for representation of a deletion in a sequence in the value for the qualifier “replace”).

Root element

43. The root element of an XML instance according to this Standard is the element ST26SequenceListing, having the following attributes:

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Mandatory/Optional</th>
</tr>
</thead>
<tbody>
<tr>
<td>dtdVersion</td>
<td>Version of the DTD used to create this file in the format “V#_#”. e.g. “V1_0”.</td>
<td>Mandatory</td>
</tr>
<tr>
<td>fileName</td>
<td>Name of the sequence listing file.</td>
<td>Optional</td>
</tr>
<tr>
<td>softwareName</td>
<td>Name of the software that generated this file.</td>
<td>Optional</td>
</tr>
<tr>
<td>softwareVersion</td>
<td>Version of the software that generated this file.</td>
<td>Optional</td>
</tr>
<tr>
<td>productionDate</td>
<td>Date of production of the sequence listing file (format “CCYY-MM-DD”).</td>
<td>Optional</td>
</tr>
</tbody>
</table>

44. The following example illustrates the root element ST26SequenceListing, and its attributes, of an XML instance as per paragraph 43 above:

```xml
<ST26SequenceListing dtdVersion="V1_0" fileName="US11_405455_SEQL.xml" softwareName="SEQL-software-name" softwareVersion="1.0" productionDate="2006-05-10"> {...}*
</ST26SequenceListing>
```

*{...}* represents the general information part and the sequence data part that have not been included in this example.
General information part

The elements of the general information part relate to patent application information, as follows:

<table>
<thead>
<tr>
<th>Element</th>
<th>Description</th>
<th>Mandatory/Optional</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ApplicationIdentification</strong></td>
<td>The application identification for which the sequence listing is submitted</td>
<td>Mandatory when a sequence listing is furnished at any time following the assignment of the application number</td>
</tr>
<tr>
<td>IPOfficeCode</td>
<td>ST.3 Code of the office of filing</td>
<td>Mandatory</td>
</tr>
<tr>
<td>ApplicationNumberText</td>
<td>The application identification as provided by the office of filing (e.g., PCT/IB2013/099999)</td>
<td>Mandatory</td>
</tr>
<tr>
<td>FilingDate</td>
<td>The date of filing of the patent application for which the sequence listing is submitted (ST.2 format “CCYY-MM-DD”, using a 4-digit calendar year, a 2-digit calendar month and a 2-digit day within the calendar month, e.g., 2015-01-31)</td>
<td>Mandatory when a sequence listing is furnished at any time following the assignment of a filing date</td>
</tr>
<tr>
<td>ApplicantFileReference</td>
<td>A single unique identifier assigned by applicant to identify a particular application, typed in the characters as set forth in paragraph 40 (b)</td>
<td>Mandatory when a sequence listing is furnished at any time prior to assignment of the application number; otherwise, Optional</td>
</tr>
<tr>
<td>EarliestPriorityApplicationIdentification</td>
<td>The application identification of the earliest priority claim (also contains IPOfficeCode, ApplicationNumberText and FilingDate, see ApplicationIdentification above)</td>
<td>Mandatory where priority is claimed</td>
</tr>
<tr>
<td>ApplicantName</td>
<td>Name of the first mentioned applicant typed in the characters as set forth in paragraph 40 (a). This element includes the mandatory attribute languageCode as set forth in paragraph 47.</td>
<td>Mandatory</td>
</tr>
<tr>
<td>ApplicantNameLatin</td>
<td>Where ApplicantName is typed in characters other than those as set forth in paragraph 40 (b), a translation or transliteration of the name of the first mentioned applicant must also be typed in characters as set forth in paragraph 40 (b)</td>
<td>Mandatory where ApplicantName contains non-Latin characters</td>
</tr>
<tr>
<td>Element</td>
<td>Description</td>
<td>Mandatory/Optional</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>InventorName</td>
<td>Name of the first mentioned inventor typed in the characters as set forth in paragraph 40 (a). This element includes the mandatory attribute languageCode as set forth in paragraph 47.</td>
<td>Optional</td>
</tr>
<tr>
<td>InventorNameLatin</td>
<td>Where InventorName is typed in characters other than those as set forth in paragraph 40 (b), a translation or transliteration of the first mentioned inventor may also be typed in characters as set forth in paragraph 40 (b)</td>
<td>Optional</td>
</tr>
<tr>
<td>InventionTitle</td>
<td>Title of the invention typed in the characters as set forth in paragraph 40 (a) in the language of filing. A translation of the title of the invention into additional languages may be typed in the characters as set forth in paragraph 40 (a) using additional InventionTitle elements. This element includes the mandatory attribute languageCode as set forth in paragraph 48. The title of invention is preferably two to seven words.</td>
<td>Mandatory in the language of filing. Optional for additional languages.</td>
</tr>
<tr>
<td>SequenceTotalQuantity</td>
<td>The total number of all sequences in the sequence listing including intentionally skipped sequences (also known as empty sequences) (see paragraph 510).</td>
<td>Mandatory</td>
</tr>
</tbody>
</table>

46. The following examples illustrate the presentation of the general information part of the sequence listing as per paragraph 45 above:

Example 1: Sequence listing filed prior to assignment of the application identification and filing date

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ST26SequenceListing PUBLIC "//WIPO/DTD Sequence Listing 1.0//EN" "ST26SequenceListing_V1_0.dtd">
<ST26SequenceListing dtdVersion="V1_0" fileName="Invention_SEQL.xml"
  softwareName="SEQL-software name" softwareVersion="1.0.0" productionDate="2015-05-10">
  <ApplicantFileReference>AB123</ApplicantFileReference>
  <EarliestPriorityApplicationIdentification>
    <IPOfficeCode>IB</IPOfficeCode>
    <ApplicationNumberText>PCT/IB2013/099999</ApplicationNumberText>
    <FilingDate>2014-07-10</FilingDate>
  </EarliestPriorityApplicationIdentification>
  <ApplicantName languageCode="EN">GENOS Co., Inc.</ApplicantName>
  <InventorName languageCode="EN">Keiko Nakamura</InventorName>
  <InventionTitle languageCode="EN">SIGNAL RECOGNITION PARTICLE RNA AND PROTEINS</InventionTitle>
  <SequenceTotalQuantity>9</SequenceTotalQuantity>
  <SequenceData sequenceIDNumber="1">{...}</SequenceData>
  <SequenceData sequenceIDNumber="2">{...}</SequenceData>
  <SequenceData sequenceIDNumber="3">{...}</SequenceData>
  <SequenceData sequenceIDNumber="4">{...}</SequenceData>
  <SequenceData sequenceIDNumber="5">{...}</SequenceData>
  <SequenceData sequenceIDNumber="6">{...}</SequenceData>
  <SequenceData sequenceIDNumber="7">{...}</SequenceData>
  <SequenceData sequenceIDNumber="8">{...}</SequenceData>
  <SequenceData sequenceIDNumber="9">{...}</SequenceData>
</ST26SequenceListing>
```
Example 2: Sequence listing filed after assignment of the application identification and filing date

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ST26SequenceListing PUBLIC "//WIPO/DTD Sequence Listing 1.0//EN" "ST26SequenceListing_V1_0.dtd">
<ST26SequenceListing dtdVersion="1.0" fileName="Invention_SEQL.xml" softwareName="SEQL-software-name" softwareVersion="1.0" productionDate="2015-05-10">
  <ApplicationIdentification>
    <IPOfficeCode>US</IPOfficeCode>
    <ApplicationNumberText>PCT/IB2014/099999</ApplicationNumberText>
    <FilingDate>2015-01-05</FilingDate>
    <RevisionSequenceList>
      <SequenceData sequenceIDNumber="1" quantity="9">...
      <SequenceData sequenceIDNumber="2" quantity="10"/>
      <SequenceData sequenceIDNumber="3" quantity="9"/>
      <SequenceData sequenceIDNumber="4" quantity="10"/>
      <SequenceData sequenceIDNumber="5" quantity="10"/>
      <SequenceData sequenceIDNumber="6" quantity="10"/>
      <SequenceData sequenceIDNumber="7" quantity="10"/>
      <SequenceData sequenceIDNumber="8" quantity="10"/>
      <SequenceData sequenceIDNumber="9" quantity="10"/>
    </RevisionSequenceList>
  </ApplicationIdentification>
  <SequenceTotalQuantity>9</SequenceTotalQuantity>
</ST26SequenceListing>
```

*{...} represents relevant information for each sequence that has not been included in this example.

47. The name of the applicant and, optionally, the name of the inventor must be indicated in the element <ApplicantName> and <InventorName>, respectively, as they are generally referred to in the language in which the application is filed. The appropriate language code (see reference in paragraph 45) must be indicated in the <languageCode> attribute for each element. Where the applicant name indicated contains characters other than those of the Latin alphabet as set forth in paragraph 40, a transliteration or translation of the applicant name must also be indicated in characters of the Latin alphabet in the element <ApplicantNameLatin>. Where the inventor name indicated contains characters other than those of the Latin alphabet, a transliteration or a translation of the inventor name may also be indicated in characters of the Latin alphabet in the element <InventorNameLatin>.

48. The title of the invention must be indicated in the element <InventionTitle> in the language of filing and may also be indicated in additional languages using multiple <InventionTitle> elements (see table in paragraph 45). The appropriate language code (see reference in paragraph 45) must be indicated in the <languageCode> attribute of the element.

49. The following example illustrates the presentation of names and title of the invention as per paragraphs 47 and 48 above:

Example: Applicant name and inventor name are each presented in Japanese and Latin characters and the title of the invention is presented in Japanese, English and French

```xml
<ApplicationName languageCode="ja">出願製薬株式会社</ApplicationName>
<ApplicationNameLatin>Shutsugan Pharmaceuticals Kabushiki Kaisha</ApplicationNameLatin>
<InventorName languageCode="ja">特許庁大郷</InventorName>
<InventorNameLatin>Taro Tokkyo</InventorNameLatin>
<InventionTitle languageCode="ja">efg タンパク質のファイルコードする手段 abcd-1 遺伝子</InventionTitle>
```
Sequence data part

50. The sequence data part must be composed of one or more SequenceData elements, each element containing information about one sequence.

51. Each SequenceData element must have a mandatory attribute sequenceIDNumber, in which the sequence identification number (see paragraph 510) for each sequence is contained. For example:

   <SequenceData sequenceIDNumber="1"/>

52. The SequenceData element must contain a dependent element INSSeq consisting of further dependent elements as follows:

<table>
<thead>
<tr>
<th>Element</th>
<th>Description</th>
<th>Mandatory/Not Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>INSSeq_length</td>
<td>Length of the sequence</td>
<td>Mandatory</td>
</tr>
<tr>
<td>INSSeq_moltype</td>
<td>Molecule type</td>
<td>Mandatory</td>
</tr>
<tr>
<td>INSSeq_division</td>
<td>Indication that a sequence is related to a patent application</td>
<td>Mandatory</td>
</tr>
<tr>
<td>INSSeq_feature-table</td>
<td>List of annotations of the sequence</td>
<td>Mandatory</td>
</tr>
<tr>
<td>INSSeq_sequence</td>
<td>Sequence</td>
<td>Mandatory</td>
</tr>
</tbody>
</table>

53. The element INSSeq_length must disclose the number of nucleotides or amino acids of the sequence contained in the INSSeq_sequence element. For example:

   <INSSeq_length>120</INSSeq_length>

54. The element INSSeq_moltype must disclose the type of molecule that is being represented. For nucleotide sequences, including nucleotide analogue sequences, the molecule type must be indicated as DNA or RNA. For protein or polypeptide amino acid sequences, the molecule type must be indicated as AA. (This element is distinct from the qualifiers “mol_type” and “MOL_TYPE” discussed in paragraphs 55 and 334). For example:

   <INSSeq_moltype>AA</INSSeq_moltype>

55. Where a nucleotide sequence contains both DNA and RNA fragments, the value for INSSeq_moltype must be “DNA.” The combined DNA/RNA molecule must be further described in the feature table, using the feature key “source” and the mandatory qualifier “organism” with the value “synthetic construct” and the mandatory qualifier “mol_type” with the value “other DNA.” Each DNA and RNA fragment of the combined DNA/RNA molecule should be further described with the feature key “misc_feature” and the qualifier “note”, which indicates whether the fragment is DNA or RNA.

56. The following example illustrates the description of a nucleotide sequence containing both DNA and RNA fragments as per paragraph 55 above:

   <INSSeq>
     <INSSeq_length>120</INSSeq_length>
     <INSSeq_moltype>DNA</INSSeq_moltype>
     <INSSeq_division>PAT</INSSeq_division>
     <INSSeq_feature-table>
       <INSDFeature>
         <INSDFeature_key>source</INSDFeature_key>
         <INSDFeature_location>1..120</INSDFeature_location>
         <INSDFeature_quals>
           <INSDFeature_quals>
             <INSDFeature_name>organism</INSDFeature_name>
57. The element `INSDSeq_sequence` must disclose the sequence. The residues in the sequence must be presented contiguously using only the appropriate symbols set forth in Annex I (see Section 1, Table 1 and Section 3, Table 3) must be included in the sequence. The sequence must not contain numbers, punctuation or whitespace characters.

58. An intentionally skipped sequence must be included in the sequence listing and represented as follows:

   (a) the element `SequenceData` and its attribute `sequenceIDNumber`, with the sequence identification number of the skipped sequence provided as the value;

   (b) the elements `INSDSeq_length`, `INSDSeq_moltype`, `INSDSeq_division`, present but with no value provided;

   (c) the element `INSDSeq_feature-table` must not be included; and

   (d) the element `INSDSeq_sequence` with the string “000” as the value.

59. The following example illustrates the representation of an intentionally skipped sequence as per paragraph 58 above:

```xml
<SequenceData sequenceIDNumber="3">
  <INSDSeq>
    <INSDSeq_length/>
    <INSDSeq_moltype/>
    <INSDSeq_division/>
    <INSDSeq_sequence>000</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
```
Feature table

60. The feature table contains information on the location and roles of various regions within a particular sequence. A feature table is required for every sequence, except for any intentionally skipped sequence, in which case it must not be included. The feature table is contained in the element `INSDSeq_feature-table`, which consists of one or more `INSDFeature` elements.

61. Each `INSDFeature` element describes one feature, and consists of dependent elements as follows:

<table>
<thead>
<tr>
<th>Element</th>
<th>Description</th>
<th>Mandatory/Optional</th>
</tr>
</thead>
<tbody>
<tr>
<td>INSDFeature_key</td>
<td>A word or abbreviation indicating a feature</td>
<td>Mandatory</td>
</tr>
<tr>
<td>INSDFeature_location</td>
<td>Region of the <code>presented</code> sequence which corresponds to the feature</td>
<td>Mandatory</td>
</tr>
<tr>
<td>INSDFeature_quals</td>
<td>Qualifier containing auxiliary information about a feature</td>
<td>Mandatory where the feature key requires one or more qualifiers, e.g., source; otherwise, Optional</td>
</tr>
</tbody>
</table>

Feature keys

62. Annex I contains an exclusive listing of feature keys that must be used under this Standard, along with an exclusive listing of associated qualifiers and an indication as to whether those qualifiers are mandatory or optional. Section 5 of Annex I provides the exclusive listing of feature keys for nucleotide sequences and Section 7 provides the exclusive listing of feature keys for amino acid sequences.

Mandatory feature keys

63. The “source” feature key is mandatory for all nucleotide sequences and the “SOURCE” feature key is mandatory for all amino acid sequences, except for any intentionally skipped sequence. Each sequence must have a single “source” or “SOURCE” feature key spanning the entire sequence. Where a sequence originates from multiple sources, those sources may be further described in the feature table, using the feature key “misc_feature” and the qualifier “note” for nucleotide sequences, and the feature key “REGION” and the qualifier “NOTE” for amino acid sequences.

64. Certain feature keys require that another feature key, referred to as a “Parent Key”, be used along with those certain feature keys. For example, the “C_region” feature key requires the “CDS” feature key (see Annex I, Section 5).

Feature location

64. The mandatory element `INSDFeature_location` must contain at least one location descriptor, which defines a site or a region corresponding to a feature of the sequence in the `INSDSeq_sequence` element, and may contain one or more location operator(s) (see paragraphs 67 to 70).

65. The location descriptor can be a single residue number, a site between two adjacent residue numbers, a region delimiting a contiguous span of residue numbers, or a site or region that extends beyond the specified residue or span of residues. Multiple location descriptors must be used in conjunction with a location operator when a feature corresponds to discontinuous sites or regions of the sequence (see paragraphs 67 to 70). The location descriptor must not include numbering for residues beyond the range of the sequence in the `INSDSeq_sequence` element.

66. The syntax for each type of location descriptor is indicated in the table below, where x and y are residue numbers, indicated as non-negative integers, not greater than the length of the sequence in the `INSDSeq_sequence` element, and x is less than y.

<table>
<thead>
<tr>
<th>Location descriptor type</th>
<th>Syntax</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single residue number</td>
<td>x</td>
<td>Points to a single residue in the <code>presented</code> sequence.</td>
</tr>
<tr>
<td>Residue numbers delimitating a sequence span</td>
<td>x..y</td>
<td>Points to a continuous range of residues bounded by and including the starting and ending residues.</td>
</tr>
<tr>
<td>Residues before the first or beyond the last specified residue number</td>
<td>&lt;x &gt;x</td>
<td>Points to a region including a specified residue or span of residues and extending beyond a specified residue. The ‘&lt;‘ and ‘&gt;’ symbols may be used with a single residue or the starting and ending residue numbers of a span of residues to indicate that a feature extends beyond the specified residue number.</td>
</tr>
</tbody>
</table>
67. A location operator is a prefix to either one location descriptor or a combination of location descriptors corresponding to a single but discontinuous feature, and specifies where the location corresponding to the feature on the indicated sequence is found or how the feature is constructed. A list of location operators is provided below with their definitions.

(a) Location operator for nucleotides and amino acids:

<table>
<thead>
<tr>
<th>Location syntax</th>
<th>Location description</th>
</tr>
</thead>
<tbody>
<tr>
<td>join(location, location, ... location)</td>
<td>The indicated locations are joined (placed end-to-end) to form one contiguous sequence.</td>
</tr>
<tr>
<td>order(location, location, ... location)</td>
<td>The elements are found in the specified order but nothing is implied about whether joining those elements is reasonable.</td>
</tr>
</tbody>
</table>

(b) Location operator for nucleotides only:

<table>
<thead>
<tr>
<th>Location syntax</th>
<th>Location description</th>
</tr>
</thead>
<tbody>
<tr>
<td>complement(location)</td>
<td>Indicates that the feature is located on the strand complementary to the sequence span specified by the location descriptor, when read in the 5' to 3' direction or in the direction that mimics the 5' to 3' direction.</td>
</tr>
</tbody>
</table>

68. The join and order location operators require that at least two comma-separated location descriptors be provided. Location descriptors involving sites between two adjacent residues, i.e. x^y, must not be used within a join or order location. Use of the join location operator implies that the residues described by the location descriptors are physically brought into contact by biological processes (for example, the exons that contribute to a coding region feature).

69. The location operator “complement” can be used for nucleotides only. “Complement” can be used in combination with either “join” or “order” within the same location. Combinations of “join” and “order” within the same location must not be used.

70. The following examples illustrate feature locations, as per paragraphs 65 to 69 above:

(a) locations for nucleotides and amino acids:

<table>
<thead>
<tr>
<th>Location Example</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>467</td>
<td>Points to residue 467 in the sequence.</td>
</tr>
<tr>
<td>123^124</td>
<td>Points to a site between residues 123 and 124.</td>
</tr>
<tr>
<td>340..565</td>
<td>Points to a continuous range of residues bounded by and including residues 340 and 565.</td>
</tr>
<tr>
<td>&lt;1</td>
<td>Points to a feature location before the first residue.</td>
</tr>
<tr>
<td>&lt;345..500</td>
<td>Indicates that the exact lower boundary point of a feature is unknown. The location begins at some residue previous to 345 and continues to and includes residue 500.</td>
</tr>
<tr>
<td>&lt;1..888</td>
<td>Indicates that the feature starts before the first sequence residue and continues to and includes residue 888.</td>
</tr>
<tr>
<td>1..&gt;888</td>
<td>Indicates that the feature starts at the first sequenced residue and continues beyond residue 888.</td>
</tr>
</tbody>
</table>
HANDBOOK ON INDUSTRIAL PROPERTY
INFORMATION AND DOCUMENTATION

Ref.: Standards - ST.26

Location Example | Description
---|---
join(12..78,134..202) | Indicates that regions 12 to 78 and 134 to 202 should be joined to form one contiguous sequence.

(b) | Location examples for nucleotides only:

<table>
<thead>
<tr>
<th>Location Example</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>complement(34..126)</td>
<td>Starts at the base nucleotide complementary to 126 and finishes at the base nucleotide complementary to base nucleotide 34 (the feature is on the strand complementary to the presented strand).</td>
</tr>
<tr>
<td>complement(join(2691..4571, 4918..5163))</td>
<td>Joins base nucleotides 2691 to 4571 and 4918 to 5163, then complements the joined segments (the feature is on the strand complementary to the presented strand).</td>
</tr>
<tr>
<td>join(complement(4918..5163), complement(2691..4571))</td>
<td>Complements regions 4918 to 5163 and 2691 to 4571, then joins the complemented segments (the feature is on the strand complementary to the presented strand).</td>
</tr>
</tbody>
</table>

Location Example

<table>
<thead>
<tr>
<th>Feature location &quot;&lt;&quot;1&quot;:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature location &quot;1..&gt;888&quot;:</td>
</tr>
</tbody>
</table>

Feature qualifiers

Qualifiers are used to supply information about features in addition to that conveyed by the feature key and feature location. There are three types of value formats to accommodate different types of information conveyed by qualifiers, namely:

(a) free text (see paragraphs 85 and 86);

(b) controlled vocabulary or enumerated values (e.g. a number or date); and

(c) sequences.

Section 6 of Annex I provides the exclusive listing of qualifiers and their specified value formats, if any, for each nucleotide feature key and Section 8 provides the exclusive listing of qualifiers for each amino acid feature key.

Any sequence encompassed by paragraph 57 which is provided as a qualifier value must be separately specified and assigned its own sequence identification number.

Mandatory feature qualifiers

One mandatory feature key, i.e., “source” for nucleotide sequences and “SOURCE” for amino acid sequences, requires two mandatory qualifiers, “organism” and “mol_type” for nucleotide sequences and “ORGANISM” and “MOL_TYPE” for amino acid sequences. Some optional feature keys also require mandatory qualifiers.

Qualifier elements

The element INSDFeature_quals contains one or more INSDQualifier elements. Each INSDQualifier element represents a single qualifier and consists of two dependent elements as follows:

<table>
<thead>
<tr>
<th>Element</th>
<th>Description</th>
<th>Mandatory/Optional</th>
</tr>
</thead>
<tbody>
<tr>
<td>INSDQualifier_name</td>
<td>Name of the qualifier (see Annex I, Sections 6 and 8)</td>
<td>Mandatory</td>
</tr>
<tr>
<td>INSDQualifier_value</td>
<td>Value of the qualifier, if any, in the specified format (see Annex I, Sections 6 and 8)</td>
<td>Mandatory, when specified (see Annex I, Sections 6 and 8)</td>
</tr>
</tbody>
</table>
77. The organism qualifier, i.e. “organism” for nucleotide sequences (see Annex I, Section 6) and “ORGANISM” for amino acid sequences (see Annex I, Section 8) must disclose the source, i.e., a single organism or origin, of the sequence that is being presented. Organism designations should be selected from a taxonomy database.

78. If the sequence is naturally occurring and the source organism has a Latin genus and species designation, that designation must be used as the qualifier value. The preferred English common name may be specified using the qualifier “note” for nucleotide sequences and the qualifier “NOTE” for amino acid sequences, but must not be used in the organism qualifier value.

79. The following examples illustrate the source of presented sequences as per paragraphs 77 and 78 above:

Example 1: Source for a nucleotide sequence

```
Example 1: Source for a nucleotide sequence

Example 2: Source for a protein amino acid sequence
```

80. If the sequence is naturally occurring and the source organism has a known Latin genus, but the species is unspecified or unidentified, then the organism qualifier value must indicate the Latin genus followed by “sp.”. For example:

```
80. If the sequence is naturally occurring and the source organism has a known Latin genus, but the species is unspecified or unidentified, then the organism qualifier value must indicate the Latin genus followed by “sp.”. For example:
```

81. If the source of the sequence is naturally occurring, but the Latin organism genus and species designation is unknown, then the organism qualifier value must be indicated as “unidentified” followed by any known taxonomic information in the qualifier “note” for nucleotide sequences and the qualifier “NOTE” for amino acid sequences. For example:

```
81. If the source of the sequence is naturally occurring, but the Latin organism genus and species designation is unknown, then the organism qualifier value must be indicated as “unidentified” followed by any known taxonomic information in the qualifier “note” for nucleotide sequences and the qualifier “NOTE” for amino acid sequences. For example:
```
82. If the sequence is naturally occurring and the source organism does not have a Latin genus and species designation, such as a virus, then another acceptable scientific name (e.g. “Canine adenovirus type 2”) must be used as the organism qualifier value. For example:

```xml
<INSDQualifier_name>organism</INSDQualifier_name>
<INSDQualifier_value>Canine adenovirus type 2</INSDQualifier_value>
```

83. If the sequence is not naturally occurring, the organism qualifier value must be indicated as “synthetic construct”. Further information with respect to the way the sequence was generated may be specified using the qualifier “note” for nucleotide sequences and the qualifier “NOTE” for amino acid sequences. For example:

```xml
<INSDSeq_feature-table>
  <INSDFeature>
    <INSDFeature_key>SOURCE</INSDFeature_key>
    <INSDFeature_location>1..40</INSDFeature_location>
    <INSDFeature_quals>
      <INSDQualifier>
        <INSDQualifier_name>ORGANISM</INSDQualifier_name>
        <INSDQualifier_value>synthetic construct</INSDQualifier_value>
      </INSDQualifier>
      <INSDQualifier>
        <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
        <INSDQualifier_value>protein</INSDQualifier_value>
      </INSDQualifier>
      <INSDQualifier>
        <INSDQualifier_name>NOTE</INSDQualifier_name>
        <INSDQualifier_value>synthetic peptide used as assay for antibodies</INSDQualifier_value>
      </INSDQualifier>
    </INSDFeature_quals>
  </INSDFeature>
</INSDSeq_feature-table>
```

84. The “mol_type” qualifier for nucleotide sequences (see Annex I, Section 6) and “MOL_TYPE” for amino acid sequences (see Annex I, Section 8) must disclose the type of molecule represented in the sequence. These qualifiers are distinct from the element INSDFeature moltype discussed in paragraph 54:

(a) For a nucleotide sequence, the “mol_type” qualifier value must be one of the following: “genomic DNA”, “genomic RNA”, “mRNA”, “rRNA”, “tRNA”, “other RNA”, “other DNA”, “transcribed RNA”, “viral cRNA”, “unassigned DNA”, or “unassigned RNA”. If the sequence is not naturally occurring, i.e. the value of the “organism” qualifier is “synthetic construct”, the “mol_type” qualifier value must be either “other RNA” or “other DNA”; (b) For an amino acid sequences, the “MOL_TYPE” qualifier value is “protein”.

Free text

85. Free text is a type of value format for certain qualifiers (as indicated in Annex I), presented in the form of a descriptive text phrase that should preferably be in the English language.

86. The use of free text must be limited to a few short terms indispensable for the understanding of a characteristic of the sequence. For each qualifier, the free text must not exceed 1000 characters.

Coding sequences

87. The “CDS” feature key may be used to identify coding sequences, i.e. sequences of nucleotides which correspond to the sequence of amino acids in a protein and the stop codon. The element INSDFeature_location should identify the location of the “CDS” feature and must include the stop codon.

88. The “transl_table” and “translation” qualifiers may be used with the “CDS” feature key (see Annex I). Where the “transl_table” qualifier is not used, the use of the Standard Code Table (see Annex I, Section 9, Table 5) is assumed.

89. The “transl_except” qualifier must be used with the “CDS” feature key and the “translation” qualifier to identify a codon that encodes either pyrrolysine or selenocysteine.
A protein amino acid sequence encoded by the coding sequence and disclosed in a “translation” qualifier that is encompassed by paragraph 57 must be included in the sequence listing and assigned its own sequence identification number and be presented in the sequence listing. The sequence identification number assigned to the protein amino acid sequence must be provided as the value in the qualifier “protein_id” with the “CDS” feature key. The “ORGANISM” qualifier of the “SOURCE” feature key for the protein amino acid sequence must be identical to that of its coding sequence. For example:

```
<INSDSeq_feature-table>
  <INSDFeature>
    <INSDFeature_key>CDS</INSDFeature_key>
    <INSDFeature_location>1..507</INSDFeature_location>
    <INSDFeature_quals>
      <INSDQualifier>
        <INSDQualifier_name>transl_table</INSDQualifier_name>
        <INSDQualifier_value>11</INSDQualifier_value>
      </INSDQualifier>
      <INSDQualifier>
        <INSDQualifier_name>translation</INSDQualifier_name>
        <INSDQualifier_value>
          MLVHLERTTMFDSSILMPLILGILAIAYLYLMGDFDLLGIGILLLPAPSDKCRDHSAISSAPWGGNHTLVLYGLGGLFAA
          FPLAYSLMPAFYIPIMLLGGLLYVLGRVPSFETRFMEKYRRLNYAFHGSILGAACFGQMIULGAFIHGVVENGRNFFGQLM
        </INSDQualifier_value>
      </INSDQualifier>
      <INSDQualifier>
        <INSDQualifier_name>protein_id</INSDQualifier_name>
        <INSDQualifier_value>89</INSDQualifier_value>
      </INSDQualifier>
    </INSDFeature_quals>
  </INSDFeature>
</INSDSeq_feature-table>
```

Variants

A primary sequence and any variant of that sequence, each disclosed by enumeration of their residues and encompassed by paragraph 57, must each be presented in the sequence listing with and assigned their own sequence identification number.

Any variant sequence, disclosed as a single sequence with enumerated alternative variant residues at one or more positions, must be included in the sequence listing and should be represented by a single sequence, wherein the enumerated alternative variant residues are represented by the most restrictive ambiguity symbol (see paragraphs 15 and 27).

Any variant sequence, disclosed only by reference to deletion(s), insertion(s), or substitution(s) in a primary sequence in the sequence listing, may be presented in the sequence listing. Where provided included in the sequence listing, such a variant sequence:

(a) may be presented by annotation of the primary sequence, where it contains variation(s) at a single location or multiple distinct locations and the occurrence of those variations are independent;

(b) should be presented as a separate sequence with and assigned its own sequence identification number, where it contains variations at multiple distinct locations and the occurrence of these variations are interdependent; and

(c) must be presented as a separate sequence with and assigned its own sequence identification number, where it contains an inserted or substituted sequence that contains in excess of 1000 residues (see paragraph 86).

The table below indicates the proper use of feature keys and qualifiers for nucleic acid and amino acid variants:

<table>
<thead>
<tr>
<th>Type of sequence</th>
<th>Feature Key</th>
<th>Qualifier</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleic acid</td>
<td>variation</td>
<td>replace or note</td>
<td>Naturally occurring mutations and polymorphisms, e.g. alleles, RFLPs.</td>
</tr>
<tr>
<td>Nucleic acid</td>
<td>misc_difference</td>
<td>replace or note</td>
<td>Variability introduced artificially, e.g. by genetic manipulation or by chemical synthesis.</td>
</tr>
</tbody>
</table>
95. Annotation of a primary sequence for a specific variant must include a feature key and qualifier, as indicated in the table above, and the feature location. The value for the “replace” qualifier must be only a single alternative nucleotide or nucleotide sequence using only the symbols in set forth Section 1, Table 1. A listing of alternative variant residues may be provided as the value in the “note” or “NOTE” qualifier. In particular, a listing of alternative amino acids must be provided as the value in the “NOTE” qualifier where “X” is used in a sequence, but represents a subgroup of “any one of A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, O, S, U, T, W, Y, or V”. A deletion must be represented by an empty qualifier value for the qualifier “replace” or an indication in the “note” or “NOTE” that the residue may be deleted. An inserted or substituted residue(s) must be provided in the “replace”, “note”, or “NOTE” qualifier. The value format for the “replace”, “note”, and “NOTE” qualifiers is free text and must not exceed 1000 characters, as provided in paragraph 86. See paragraph 96 for sequences encompassed by paragraph 37 that are provided as an insertion or a substitution in a qualifier value. A listing of alternative residues for an insertion or substitution may be provided as the qualifier value.

96. The symbols set forth in Annex I (see Sections 1 to 4, Tables 1 to 4, respectively) should be used to represent variant residues where appropriate. Where the variant residue is a modified residue not set forth in Tables 2 or 4 of Annex I, the complete unabbreviated name of the modified residue must be provided as the qualifier value. Modified residues must be further described in the feature table as provided in paragraph 17 or 30.

97. The following examples illustrate the presentation of variants as per paragraphs 92 to 95 above:

Example 1: Feature key “variation_mis_difference” for a substitution in a nucleotide sequence enumerated alternative variant nucleotides. The “n” at position 53 of the sequence can be one of five alternative nucleotides: A cytosine replaces the nucleotide given in position 413 of the sequence.

Example 2: Feature key “misc_difference” for a deletion in a nucleotide sequence. The nucleotide at position 413 of the sequence is deleted.

<table>
<thead>
<tr>
<th>Type of sequence</th>
<th>Feature Key</th>
<th>Qualifier</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acid</td>
<td>VAR_SEQ</td>
<td>NOTE</td>
<td>Variant produced by alternative splicing, alternative promoter usage, alternative initiation and ribosomal frameshifting.</td>
</tr>
<tr>
<td>Amino acid</td>
<td>VARIANT</td>
<td>NOTE</td>
<td>Any type of variant for which VAR_SEQ is not applicable.</td>
</tr>
</tbody>
</table>
Example 3: Feature key “misc_difference” for an insertion in a nucleotide sequence. The sequence “atgccaaatat” is inserted between positions 100 and 101 of the primary sequence.

Example 4: Feature key “VARIANT” variation for a substitution in an amino acid, a nucleotide sequence.

Example 5: Feature key “VARIANT” for a substitution in an amino acid sequence. The amino acid given in position 100 of the sequence can be replaced by any amino acid except for Lys, Arg or His.

Example 6: Feature key “VARIANT” for a substitution in an amino acid sequence. The amino acid given in position 100 of the sequence can be replaced by any amino acid except for Lys, Arg or His.
<INSDQualifier_value>not K, R, or H</INSDQualifier_value>
</INSDFeature quoi>  
</INSDFeature>

98. A sequence encompassed by paragraph 37 that is provided as an insertion or a substitution in a qualifier value for a primary sequence annotation must also be included in the sequence listing and assigned its own sequence identification number.

[Annex I follows]
ANNEX I

CONTROLLED VOCABULARY

Version 1.01.1

Adopted and Approved by the Committee on WIPO Standards (CWS)
at its reconvened fourth session on March 24, 2016 and fifth session on June 2, 2017

TABLE OF CONTENTS

SECTION 1: LIST OF NUCLEOTIDES ................................................................. 2
SECTION 2: LIST OF MODIFIED NUCLEOTIDES ......................................... 2
SECTION 3: LIST OF AMINO ACIDS ............................................................... 4
SECTION 4: LIST OF MODIFIED AND UNUSUAL AMINO ACIDS ...................... 5
SECTION 5: FEATURE KEYS FOR NUCLEIC ACID SEQUENCES .................. 6
SECTION 6: DESCRIPTION OF QUALIFIERS FOR NUCLEIC ACID SEQUENCES . 27
SECTION 7: FEATURE KEYS FOR AMINO ACID SEQUENCES ....................... 50
SECTION 8: QUALIFIERS FOR AMINO ACID SEQUENCES ............................. 57
SECTION 9: GENETIC CODE TABLES ............................................................... 58
SECTION 1: LIST OF NUCLEOTIDES
The nucleotide base codes to be used in sequence listings are presented in Table 1. The symbol “t” will be construed as thymine in DNA and uracil in RNA when it is used with no further description. Where an ambiguity symbol (representing two or more bases in the alternative) is appropriate, the most restrictive symbol should be used. For example, if a base in a given position could be “a or g,” then “r” should be used, rather than “n”. The symbol “n” will be construed as “a or c or g or t/u” when it is used with no further description.

Table 1: List of nucleotides

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Nucleotide</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>adenine</td>
</tr>
<tr>
<td>c</td>
<td>cytosine</td>
</tr>
<tr>
<td>g</td>
<td>guanine</td>
</tr>
<tr>
<td>t</td>
<td>thymine in DNA/uracil in RNA (t/u)</td>
</tr>
<tr>
<td>m</td>
<td>a or c</td>
</tr>
<tr>
<td>r</td>
<td>a or g</td>
</tr>
<tr>
<td>w</td>
<td>a or t/u</td>
</tr>
<tr>
<td>s</td>
<td>c or g</td>
</tr>
<tr>
<td>y</td>
<td>c or t/u</td>
</tr>
<tr>
<td>k</td>
<td>g or t/u</td>
</tr>
<tr>
<td>v</td>
<td>a or c or g; not t/u</td>
</tr>
<tr>
<td>h</td>
<td>a or c or t/u; not g</td>
</tr>
<tr>
<td>d</td>
<td>a or g or t/u; not c</td>
</tr>
<tr>
<td>b</td>
<td>c or g or t/u; not a</td>
</tr>
<tr>
<td>n</td>
<td>a or c or g or t/u; “unknown” or “other”</td>
</tr>
</tbody>
</table>

SECTION 2: LIST OF MODIFIED NUCLEOTIDES
The abbreviations listed in Table 2 are the only permitted values for the mod_base qualifier. Where a specific modified nucleotide is not present in the table below, then the abbreviation “OTHER” must be used as its value. If the abbreviation is “OTHER,” then the complete unabbreviated name of the modified base must be provided in a note qualifier. The abbreviations provided in Table 2 must not be used in the sequence itself.

Table 2: List of modified nucleotides

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Modified Nucleotide</th>
</tr>
</thead>
<tbody>
<tr>
<td>ac4c</td>
<td>4-acetylcytidine</td>
</tr>
<tr>
<td>chm5u</td>
<td>5-(carboxyhydroxymethyl)uridine</td>
</tr>
<tr>
<td>cm</td>
<td>2’-O-methylcytidine</td>
</tr>
<tr>
<td>cmnm5s2u</td>
<td>5-carboxymethylaminomethyl-2-thiouridine</td>
</tr>
<tr>
<td>cmnm5u</td>
<td>5-carboxymethylaminomethyluridine</td>
</tr>
<tr>
<td>3dhu</td>
<td>dihydrouridine</td>
</tr>
<tr>
<td>fm</td>
<td>2’-O-methylpseudouridine</td>
</tr>
<tr>
<td>gal q</td>
<td>beta-D-\text{galactosylqueosine} \text{galactosylqueuosine}</td>
</tr>
<tr>
<td>gm</td>
<td>2’-O-methylguanosine</td>
</tr>
<tr>
<td>i</td>
<td>inosine</td>
</tr>
<tr>
<td>i6a</td>
<td>N6-isopentenyladenosine</td>
</tr>
<tr>
<td>m1a</td>
<td>1-methyladenosine</td>
</tr>
<tr>
<td>m1f</td>
<td>1-methylpseudouridine</td>
</tr>
<tr>
<td>m1g</td>
<td>1-methylguanosine</td>
</tr>
<tr>
<td>m1i</td>
<td>1-methylinosine</td>
</tr>
<tr>
<td>m22g</td>
<td>2,2-dimethylguanosine</td>
</tr>
<tr>
<td>m2a</td>
<td>2-methyladenosine</td>
</tr>
<tr>
<td>m2g</td>
<td>2-methylguanosine</td>
</tr>
<tr>
<td>m3c</td>
<td>3-methylcytidine</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Modified Nucleotide</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>m4c</td>
<td>N4-methylcytosine</td>
</tr>
<tr>
<td>m5c</td>
<td>5-methylcytidine</td>
</tr>
<tr>
<td>m6a</td>
<td>N6-methyladenosine</td>
</tr>
<tr>
<td>m7g</td>
<td>7-methylguanosine</td>
</tr>
<tr>
<td>mam5u</td>
<td>5-methylaminomethyluridine</td>
</tr>
<tr>
<td>mam5s2u</td>
<td>5'-methylaminomethylmethylaminomethyl-2-thiouridine</td>
</tr>
<tr>
<td>man q</td>
<td>beta-D-mannosylqueosine</td>
</tr>
<tr>
<td>mcm5s2u</td>
<td>5-methoxy carbonylmethyl-2-thiouridine</td>
</tr>
<tr>
<td>mcm5u</td>
<td>5-methoxy carbonylmethyluridine</td>
</tr>
<tr>
<td>mo5u</td>
<td>5-methoxyuridine</td>
</tr>
<tr>
<td>ms2i6a</td>
<td>2-methythio-N6-isopentenyladenosine</td>
</tr>
<tr>
<td>ms2t6a</td>
<td>N-((9-beta-D-ribofuranosyl-2'-methylthiopurine)methylthiopurine-6-yl)carbamoylthreonine</td>
</tr>
<tr>
<td>mt6a</td>
<td>N-((9-beta-D-ribofuranosylpurine-6-yl)N-methyl-carbamoyl)threonine</td>
</tr>
<tr>
<td>mv</td>
<td>uridine-5'-oxoacetic acid-methylester</td>
</tr>
<tr>
<td>o5u</td>
<td>uridine-5'-oxyacetic acid (v)</td>
</tr>
<tr>
<td>osyw</td>
<td>wybutoxosine</td>
</tr>
<tr>
<td>p</td>
<td>pseudouridine</td>
</tr>
<tr>
<td>q</td>
<td>queosine</td>
</tr>
<tr>
<td>s2c</td>
<td>2-thiocytidine</td>
</tr>
<tr>
<td>s2t</td>
<td>5-methyl-2-thiouridine</td>
</tr>
<tr>
<td>s2u</td>
<td>2-thiouridine</td>
</tr>
<tr>
<td>s4u</td>
<td>4-thiouridine</td>
</tr>
<tr>
<td>m5u</td>
<td>5-methyluridine</td>
</tr>
<tr>
<td>t6a</td>
<td>N-((9-beta-D-ribofuranosylpurine-6-yl)carbamoyl)threonine</td>
</tr>
<tr>
<td>tm</td>
<td>2'-O-methyl-5-methyluridine</td>
</tr>
<tr>
<td>um</td>
<td>2'-O-methyluridine</td>
</tr>
<tr>
<td>yw</td>
<td>wybutosine</td>
</tr>
<tr>
<td>x</td>
<td>3-(3-amino-3-carboxypropyl)uridine, (acp3)u</td>
</tr>
<tr>
<td>OTHER</td>
<td>(requires note qualifier)</td>
</tr>
</tbody>
</table>
SECTION 3: LIST OF AMINO ACIDS

The amino acid codes to be used in sequence listings are presented in Table 3. Where an ambiguity symbol (representing two or more amino acids in the alternative) is appropriate, the most restrictive symbol should be used. For example, if an amino acid in a given position could be aspartic acid or asparagine, the symbol “B” should be used, rather than “X”. The symbol “X” will be construed as one of “A”, “R”, “N”, “D”, “C”, “Q”, “E”, “G”, “H”, “I”, “L”, “K”, “M”, “F”, “P”, “O”, “S”, “U”, “T”, “W”, “Y”, or “V”, when it is used with no further description.

Table 3: List of amino acids

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Amino acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Alanine</td>
</tr>
<tr>
<td>R</td>
<td>Arginine</td>
</tr>
<tr>
<td>N</td>
<td>Asparagine</td>
</tr>
<tr>
<td>D</td>
<td>Aspartic acid (Aspartate)</td>
</tr>
<tr>
<td>C</td>
<td>Cysteine</td>
</tr>
<tr>
<td>Q</td>
<td>Glutamine</td>
</tr>
<tr>
<td>E</td>
<td>Glutamic acid (Glutamate)</td>
</tr>
<tr>
<td>G</td>
<td>Glycine</td>
</tr>
<tr>
<td>H</td>
<td>Histidine</td>
</tr>
<tr>
<td>I</td>
<td>Isoleucine</td>
</tr>
<tr>
<td>L</td>
<td>Leucine</td>
</tr>
<tr>
<td>K</td>
<td>Lysine</td>
</tr>
<tr>
<td>M</td>
<td>Methionine</td>
</tr>
<tr>
<td>F</td>
<td>Phenylalanine</td>
</tr>
<tr>
<td>P</td>
<td>Proline</td>
</tr>
<tr>
<td>O</td>
<td>Pyrolysine</td>
</tr>
<tr>
<td>S</td>
<td>Serine</td>
</tr>
<tr>
<td>U</td>
<td>Selenocysteine</td>
</tr>
<tr>
<td>T</td>
<td>Threonine</td>
</tr>
<tr>
<td>W</td>
<td>Tryptophan</td>
</tr>
<tr>
<td>Y</td>
<td>Tyrosine</td>
</tr>
<tr>
<td>V</td>
<td>Valine</td>
</tr>
<tr>
<td>B</td>
<td>Aspartic acid or Asparagine</td>
</tr>
<tr>
<td>Z</td>
<td>Glutamine or Glutamic acid</td>
</tr>
<tr>
<td>J</td>
<td>Leucine or Isoleucine</td>
</tr>
<tr>
<td>X</td>
<td>unknown or other: A or R or N or D or C or Q or E or G or H or I or L or K or M or F or P or O or S or U or T or W or Y or V; “unknown” or “other”</td>
</tr>
</tbody>
</table>
SECTION 4: LIST OF MODIFIED AND UNUSUAL AMINO ACIDS

Table 4 lists the only permitted abbreviations for a modified or unusual amino acid in the mandatory qualifier “NOTE” for feature keys “MOD_RES” or “SITE”. The value for the qualifier “NOTE” must be either an abbreviation from this table, where appropriate, or the complete, unabbreviated name of the modified amino acid. The abbreviations (or full names) provided in this table must not be used in the sequence itself.

Table 4: List of modified and unusual amino acids

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Modified or Unusual Amino acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aad</td>
<td>2-Aminoacidipic acid</td>
</tr>
<tr>
<td>bAad</td>
<td>3-Aminoacidipic acid</td>
</tr>
<tr>
<td>bAla</td>
<td>beta-Alanine, beta-Aminoproprionic acid</td>
</tr>
<tr>
<td>Abu</td>
<td>2-Aminobutyric acid</td>
</tr>
<tr>
<td>4Abu</td>
<td>4-Aminobutyric acid, piperidinic acid</td>
</tr>
<tr>
<td>Acp</td>
<td>6-Aminocaproic acid</td>
</tr>
<tr>
<td>Ahe</td>
<td>2-Aminoheptanoic acid</td>
</tr>
<tr>
<td>Alb</td>
<td>2-Aminoisobutyric acid</td>
</tr>
<tr>
<td>bAlb</td>
<td>3-Aminoisobutyric acid</td>
</tr>
<tr>
<td>Apm</td>
<td>2-Aminopimelic acid</td>
</tr>
<tr>
<td>Dbu</td>
<td>2,4-Diaminobutyric acid</td>
</tr>
<tr>
<td>Des</td>
<td>Desmosine</td>
</tr>
<tr>
<td>Dpm</td>
<td>2,2'-Diaminopimelic acid</td>
</tr>
<tr>
<td>Dpr</td>
<td>2,3-Diaminopropionic acid</td>
</tr>
<tr>
<td>EtGly</td>
<td>N-Ethylglycine</td>
</tr>
<tr>
<td>EtAsn</td>
<td>N-Ethylasparagine</td>
</tr>
<tr>
<td>Hyl</td>
<td>Hydroxylysine</td>
</tr>
<tr>
<td>aHyl</td>
<td>allo-Hydroxylysine</td>
</tr>
<tr>
<td>3Hyp</td>
<td>3-Hydroxyproline</td>
</tr>
<tr>
<td>4Hyp</td>
<td>4-Hydroxyproline</td>
</tr>
<tr>
<td>Ide</td>
<td>Isodesmosine</td>
</tr>
<tr>
<td>aIle</td>
<td>allo-Isoleucine</td>
</tr>
<tr>
<td>MeGly</td>
<td>N-Methylglycine, sarcosine</td>
</tr>
<tr>
<td>MeIle</td>
<td>N-Methylisoleucine</td>
</tr>
<tr>
<td>MeLys</td>
<td>6-N-Methyllysine</td>
</tr>
<tr>
<td>MeVal</td>
<td>N-Methylvaline</td>
</tr>
<tr>
<td>Nva</td>
<td>Norvaline</td>
</tr>
<tr>
<td>Nle</td>
<td>Norleucine</td>
</tr>
<tr>
<td>Orn</td>
<td>Ornithine</td>
</tr>
</tbody>
</table>

Date: June 2016, September 2017
### SECTION 5: FEATURE KEYS FOR NUCLEIC ACID SEQUENCES

This section contains the list of allowed feature keys to be used for nucleic acid sequences, and lists mandatory and optional qualifiers. The feature keys are listed in alphabetic order. The feature keys can be used for either DNA or RNA unless otherwise indicated under “Molecule scope”. Some feature keys include a ‘Parent Key’ designation; in the description of a feature key, a mandatory that the designated parent key be used. Certain Feature Keys may be appropriate for use with artificial sequences in addition to the specified “organism scope”.

Feature key names must be used in the XML instance of the sequence listing exactly as they appear following “Feature key” in the descriptions below, except for the feature keys 3'UTR and 5'UTR. See “Comment” in the description for the 3'UTR and 5'UTR feature keys.

<table>
<thead>
<tr>
<th>Feature Key</th>
<th>Definition</th>
<th>Organism scope</th>
<th>Molecule scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>attenuator</td>
<td>1) region of DNA at which regulation of termination of transcription occurs, which controls the expression of some bacterial operons; sequence segment located between the promoter and the first structural gene that causes partial termination of transcription</td>
<td>prokaryotes</td>
<td>DNA</td>
</tr>
<tr>
<td>C_region</td>
<td>constant region of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; includes one or more exons depending on the particular chain</td>
<td>eukaryotes</td>
<td>DNA</td>
</tr>
<tr>
<td>CAAT_signal</td>
<td>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</td>
<td>prokaryotes</td>
<td>DNA</td>
</tr>
</tbody>
</table>
Organism scope: eukaryotes and eukaryotic viruses

Molecule scope: DNA

References:

5.2. Feature Key: CDS

Definition: coding sequence; sequence of nucleotides that corresponds with the sequence of amino acids in a protein (location includes stop codon); feature may include amino acid conceptual translation

Optional qualifiers:
- allele
- artificial_location
- codon_start
- EC_number
- exception
- function
- gene
- gene_synonym
- map
- note
- operon
- product
- protein_id
- pseudo
- pseudogene
- ribosomal_slippage
- standard_name
- translation
- transl_except
- transl_table
- trans_splicing

Comment: codon_start qualifier has valid value of 1 or 2 or 3, indicating the offset at which the first complete codon of a coding feature can be found, relative to the first base of that feature; transl_table defines the genetic code table used if other than the Standard or universal genetic code table; genetic code exceptions outside the range of the specified tables are reported in transl_except qualifier; only one of the qualifiers translation, pseudo, pseudogene or pseudo are permitted with a CDS feature key; when the translation qualifier is used, the protein_id qualifier is mandatory if the translation product contains four or more specifically defined amino acids

5.3. Feature Key: centromere

Definition: region of biological interest identified as a centromere and which has been experimentally characterized

Optional qualifiers:
- note
- standard_name

Comment: the centromere feature describes the interval of DNA that corresponds to a region where chromatids are held and a kinetochore is formed

Ref.: Standards - ST.26
Date: June 2016 September 2017
5.4. 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 allele
gene
gene_synonym
map
note

Molecule scope DNA

5.5. Feature Key D_segment

Definition Diversity segment of immunoglobulin heavy chain, and T-cell receptor beta chain

Optional qualifiers allele
gene
gene_synonym
map
note
product
pseudo
pseudogene
standard_name

Organism scope eukaryotes

Organism scope eukaryotes and eukaryotic viruses

5.6. Feature Key exon

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

Optional qualifiers allele
EC_number
function
gene
gene_synonym
map
note
number
product
pseudo
pseudogene
### 5.10. Feature Key: trans_splicing

**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**
- allele
- gene
- gene_synonym
- map
- note
- operon
- product
- pseudo
- pseudogene
- phenotype
- standard_name
- trans_splicing

**Comment**
The gene feature describes the interval of DNA that corresponds to a genetic trait or phenotype; the feature is, by definition, not strictly bound to its positions at the ends; it is meant to represent a region where the gene is located.

### 5.7. Feature Key: gene

**Definition**
region of biological interest identified as a gene and for which a name has been assigned

**Optional qualifiers**
- allele
- function
- gene
- gene_synonym
- map
- note
- operon
- product
- pseudo
- pseudogene
- phenotype
- standard_name
- trans_splicing

**Comment**
The gene feature describes the interval of DNA that corresponds to a genetic trait or phenotype; the feature is, by definition, not strictly bound to its positions at the ends; it is meant to represent a region where the gene is located.

### 5.8. Feature Key: iDNA

**Definition**
intervening DNA; DNA which is eliminated through any of several kinds of recombination

**Optional qualifiers**
- allele
- function
- gene
- gene_synonym
- map
- note
- number
- standard_name

**Molecule scope**
DNA

**Comment**
E.g., in the somatic processing of immunoglobulin genes.
5.9. 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  allele
                      function
                      gene
                      gene_synonym
                      map
                      note
                      number
                      pseudo
                      pseudogene
                      standard_name
                      trans_splicing

5.10. Feature Key  J_segment

Definition  joining segment of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains

Optional qualifiers  allele
                      gene
                      gene_synonym
                      map
                      note
                      product
                      pseudo
                      pseudogene
                      standard_name

Organism scope  eukaryotes

5.15. 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  allele
                      gene
                      gene_synonym
                      map
                      note
                      standard_name
5.11. 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  allele
EC_number
function
gene
gene_synonym
map
note
product
pseudo
pseudogene
standard_name

5.12. 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

Optional qualifiers  allele
function
gene
gene_synonym
map
note

Comment  note that the regulatory feature key and regulatory class qualifier with the value "ribosome_binding_site" must be used for describing ribosome binding sites

5.13. Feature Key  misc_difference

Definition  featured sequence differs from the presented sequence at this location and cannot be described by any other Difference key (unsure, variation, or modified_base)

Optional qualifiers  allele
clone
compare
gene
gene_synonym
map
note
phenotype
replace
standard_name

Comment  the misc_difference feature key must be used to describe variability introduced artificially, e.g. by genetic manipulation or by chemical synthesis; use the replace qualifier to annotate a deletion, insertion, or substitution. The variation feature key must be used to describe naturally occurring genetic variability.
5.14. 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** | allele, function, gene, gene_synonym, map, note, number, phenotype, product, pseudo, pseudogene, standard_name

**Comment** | 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

5.15. 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 or qualifiers of source key (proviral)

**Optional qualifiers** | allele, gene, gene_synonym, map, note, recombination_class, standard_name

**Molecule scope** | DNA

5.16. Feature Key | misc_RNA
---|---
**Definition** | any transcript or RNA product that cannot be defined by other RNA keys (prim_transcript, precursor_RNA, mRNA, 5'UTR, 3'UTR, exon, CDS, sig_peptide, transit_peptide, mat_peptide, intron, polyA_site, ncRNA, rRNA and tRNA)

**Optional qualifiers** | allele, function, gene, gene_synonym, map, note, operon, product, pseudo, pseudogene, standard_name, trans_splicing

5.22. 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, rep_origin)
<table>
<thead>
<tr>
<th>Feature Key</th>
<th>Definition</th>
<th>Optional qualifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>misc_structure</td>
<td>any secondary or tertiary nucleotide structure or conformation that cannot be described by other Structure keys (stem_loop and D-loop)</td>
<td>allele, function, gene, gene_synonym, map, note, standard_name</td>
</tr>
<tr>
<td>mobile_element</td>
<td>region of genome containing mobile elements</td>
<td>mobile_element_type, allele, function, gene, gene_synonym, map, note, rpt_family, rpt_type, standard_name</td>
</tr>
<tr>
<td>modified_base</td>
<td>the indicated nucleotide is a modified nucleotide and should be substituted for by the indicated molecule (given in the mod_base qualifier value)</td>
<td>mod_base, allele, frequency, gene, gene_synonym, map, note</td>
</tr>
</tbody>
</table>

Comment: value for the mandatory mod_base qualifier is limited to the restricted vocabulary for modified base abbreviations in Section 2 of this Annex.
5.20. Feature Key mRNA

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

Optional qualifiers allele

5.21. Feature Key ncRNA

Definition a non-protein-coding gene, other than ribosomal RNA and transfer RNA, the functional molecule of which is the RNA transcript

Mandatory qualifiers ncRNA_class

Optional qualifiers allele

Comment the ncRNA feature must not be used for ribosomal and transfer RNA annotation, for which the rRNA and tRNA feature keys must be used, respectively

5.22. Feature Key N_region

Definition extra nucleotides inserted between rearranged immunoglobulin segments

Optional qualifiers allele

Organism scope eukaryotes
5.23. Feature Key: operon

**Definition:** region containing polycistronic transcript including a cluster of genes that are under the control of the same regulatory sequences/promoter and in the same biological pathway

**Mandatory qualifiers:** operon

**Optional qualifiers:** allele, function, map, note, phenotype, pseudo, pseudogene, standard_name

5.24. Feature Key: oriT

**Definition:** origin of transfer; region of a DNA molecule where transfer is initiated during the process of conjugation or mobilization

**Optional qualifiers:** allele, bound_moiety, direction, gene, gene_synonym, map, note, rpt_family, rpt_type, rpt_unit_range, rpt_unit_seq, standard_name

**Molecule Scope:** DNA

**Comment:** rep_origin must be used to describe origins of replication; direction qualifier has legal values left, right, and both, however only left and right are valid when used in conjunction with the oriT feature; origins of transfer can be present in the chromosome; plasmids can contain multiple origins of transfer

5.31. 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:** allele, gene, gene_synonym, map, note

**Organism scope:** eukaryotes and eukaryotic viruses

5.25. Feature Key  polyA_site
Definition  site on an RNA transcript to which will be added adenine residues by post-transcriptional polyadenylation
Optional qualifiers  allele
                               gene
                               gene_synonym
                               map
                               note
Organism scope  eukaryotes and eukaryotic viruses

5.26. Feature Key  precursor_RNA
Definition  any RNA species that is not yet the mature RNA product; may include ncRNA, rRNA, tRNA, 5' untranslated region (5'UTR), coding sequences (CDS, exon), intervening sequences (intron) and 3' untranslated region (3'UTR)
Optional qualifiers  allele
                                   function
                                   gene
                                   gene_synonym
                                   map
                                   note
                                   operon
                                   product
                                   standard_name
                                   trans_splicing
Comment  used for RNA which may be the result of post-transcriptional processing; if the RNA in question is known not to have been processed, use the prim_transcript key

5.27. Feature Key  prim_transcript
Definition  primary (initial, unprocessed) transcript; may include ncRNA, tRNA, 5' untranslated region (5'UTR), coding sequences (CDS, exon), intervening sequences (intron) and 3' untranslated region (3'UTR)
Optional qualifiers  allele
                                      gene
                                      gene_synonym
                                      map
                                      note
                                      operon
                                      standard_name
Feature Key primer_bound

Definition non-covalent primer binding site for initiation of replication, transcription, or reverse transcription; includes site(s) for synthetic e.g., PCR primer elements.

Optional qualifiers:
- allele
- gene
- gene_synonym
- map
- note
- standard_name
- PCR_conditions

Comment:
used to annotate the site on a given sequence to which a primer molecule binds - not intended to represent the sequence of the primer molecule itself; PCR components and reaction settings may be stored under the PCR_conditions qualifier; since PCR reactions most often involve pairs of primers, a single primer_bound key may use the order(location, location) operator with two locations, or a pair of primer_bound keys may be used.

Feature Key protein_bound

Definition non-covalent protein binding site on nucleic acid

Mandatory qualifiers:
- bound_moiety

Optional qualifiers:
- allele
- function
- gene
- gene_synonym
- map
- note
- operon
- standard_name

Comment:
note that the regulatory feature key and regulatory_class qualifier with the value "ribosome_binding_site" must be used to describe ribosome binding sites.
5.31. Feature Key \textit{\texttt{ribosome\_binding\_site}}

\textbf{Definition:} ribosome binding site

\textbf{Optional Definition:} any region of a sequence that functions in the regulation of translation, replication or chromatin structure.

\textbf{Mandatory Qualifiers:} regulatory_class, gene_synonym, map, note, pseudo, pseudogene, standard_name


\textbf{Comment:} in prokaryotes, known as the Shine-Dalgarno sequence; is located 5 to 9 bases upstream of the initiation codon; consensus context [GGA].

5.32. Feature Key \textit{\texttt{repeat\_region}}

\textbf{Definition:} region of genome containing repeating units

\textbf{Optional Qualifiers:} allele, function, gene, gene_synonym, map, note, rpt_family, rpt_type, rpt_unit_range, rpt_unit_seq, satellite, standard_name

5.33. Feature Key \textit{\texttt{rep\_origin}}

\textbf{Definition:} origin of replication; starting site for duplication of nucleic acid to give two identical copies

\textbf{Optional Qualifiers:} allele, direction

\textbf{Comment:} direction qualifier has valid values: \texttt{LEFT, RIGHT, BOTH}. 

\hspace{1cm}
### Feature Key: rRNA

**Definition:** mature ribosomal RNA; RNA component of the ribonucleoprotein particle (ribosome) which assembles amino acids into proteins  

**Optional qualifiers:** allele, function, gene, gene_synonym, map, note, operon, product, pseudo, pseudogene, standard_name  

**Comment:** rRNA sizes should be annotated with the product qualifier

### Feature Key: S_region

**Definition:** switch region of immunoglobulin heavy chains; involved in the rearrangement of heavy chain DNA leading to the expression of a different immunoglobulin class from the same B-cell  

**Optional qualifiers:** allele, gene, gene_synonym, map, note, product, pseudo, pseudogene, standard_name

**Parent Key:** misc_signal  
**Organism scope:** eukaryotes

### 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:** allele, function, gene, gene_synonym, map, note, product, pseudo, pseudogene, standard_name
### 5.37. Feature Key

**source**

**Definition**
identifies the source of the sequence; this key is mandatory; every sequence will have a single source key spanning the entire sequence.

**Mandatory qualifiers**
- organism
- mol_type

**Optional qualifiers**
- cell_line
- cell_type
- chromosome
- clone
- clone_lib
- collected_by
- collection_date
- cultivar
- dev_stage
- ecotype
- environmental_sample
- germline
- haplogroup
- haplotype
- host
- identified_by
- isolate
- isolation_source
- lab_host
- lat_lon
- macronuclear
- map
- mating_type
- note
- organelle
- PCR_primers
- plasmid
- pop_variant
- proviral
- rearranged
- segment
- serotype
- serovar
- sex
- strain
- sub_clone
- sub_species
- sub_strain
- tissue_lib
- tissue_type
- variety

**Molecule scope**
any

### 5.38. 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**
- allele
- function
- gene
- gene_synonym
- map
- note
- operon
- standard_name
5.39. Feature Key

STS

Definition
sequence tagged site; short, single-copy DNA sequence that characterizes a mapping landmark on the genome and can be detected by PCR; a region of the genome can be mapped by determining the order of a series of STSs

Optional qualifiers
allele
gene
gene_synonym
map
note
standard_name

Molecule scope
DNA

Parent key
misc_binding

Comment
STS location to include primer(s) in primer_bind key or primers

5.40. Feature Key
telomere

Definition
region of biological interest identified as a telomere and which has been experimentally characterized

Optional qualifiers
note
rpt_type
rpt_unit_range
rpt_unit_seq
standard_name

Comment
the telomere feature describes the interval of DNA that corresponds to a specific structure at the end of the linear eukaryotic chromosome which is required for the integrity and maintenance of the end; this region is unique compared to the rest of the chromosome and represents the physical end of the chromosome
### 5.49. Feature Key: terminator

**Definition:** sequence of DNA located either at the end of the transcript that causes RNA polymerase to terminate transcription

**Optional qualifiers:**
- allele
- gene
- gene_synonym
- map
- note
- operon
- standard_name

### 5.41. Feature Key: tmRNA

**Definition:** transfer messenger RNA; tmRNA acts as a tRNA first, and then as an mRNA that encodes a peptide tag; the ribosome translates this mRNA region of tmRNA and attaches the encoded peptide tag to the C-terminus of the unfinished protein; this attached tag targets the protein for destruction or proteolysis

**Optional qualifiers:**
- allele
- function
- gene
- gene_synonym
- map
- note
- product
- pseudo
- pseudogene
- standard_name
- tag_peptide

### 5.42. Feature Key: transit_peptide

**Definition:** transit peptide coding sequence; coding sequence for an N-terminal domain of a nuclear-encoded organelar protein; this domain is involved in post-translational import of the protein into the organelle

**Optional qualifiers:**
- allele
- function
- gene
- gene_synonym
- map
- note
- product
- pseudo
- pseudogene
- standard_name
5.43. 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 | allele, anticodon, function, gene, gene_synonym, map, note, product, pseudo, pseudogene, standard_name, trans_splicing

5.44. Feature Key | unsure
--- | ---
Definition | a small region of sequenced bases, generally 10 or fewer in its length, which could not be confidently identified. Such a region might contain called bases (a, t, g, or c), or a mixture of called-bases and un-called-bases ('n').
Optional qualifiers | allele, compare, gene, gene_synonym, map, note, replace
Comment | use the replace qualifier to annotate a deletion, insertion, or substitution.

5.45. Feature Key | V_region
--- | ---
Definition | variable region of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; codes for the variable amino terminal portion; can be composed of V_segments, D_segments, N_regions, and J_segments
Optional qualifiers | allele, gene, gene_synonym, map, note, product, pseudo, pseudogene, standard_name
Parent Key | CDS
Organism scope | eukaryotes
5.46. Feature Key  
V_segment

Definition
variable segment of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; codes for most of the variable region (V_region) and the last few amino acids of the leader peptide

Optional qualifiers
allele  
gene  
gene_synonym  
map  
ote  
product  
pseudo  
pseudogene  
standard_name

Parent Key: CDS

Organism scope: eukaryotes

5.47. 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
allele  
compare  
frequency  
gene  
gene_synonym  
map  
ote  
phenotype  
product  
replace  
standard_name

Comment
used to describe alleles, RFLP’s, and other naturally occurring mutations and polymorphisms: use the replace qualifier to annotate a deletion, insertion, or substitution; variability arising as a result of genetic manipulation (e.g. site directed mutagenesis) must be described with the misc_difference feature

5.48. Feature Key  
3'UTR

Definition
1) region at the 3' end of a mature transcript (following the stop codon) that is not translated into a protein;  
2) region at the 3' end of an RNA virus (following the last stop codon) that is not translated into a protein;

Optional qualifiers
allele  
function  
gene  
gene_synonym  
map  
ote  
standard_name  
trans_splicing

Comment
The apostrophe character has special meaning in XML, and must be substituted with “&apos;” in the value of an element. Thus “3’UTR” must be represented as “3&amp;apos;UTR” in the XML file, i.e., &lt;INSDFeature_key&gt;3&amp;apos;UTR&lt;/INSDFeature_key&gt;.
### Feature Key: 5'UTR

**Definition**

1. Region at the 5' end of a mature transcript (preceding the initiation codon) that is not translated into a protein;
2. Region at the 5' end of an RNA virus (preceding the first initiation codon) that is not translated into a protein;

**Optional qualifiers**

- allele
- function
- gene
- gene_synonym
- map
- note
- standard_name
- trans_splicing

**Comment**
The apostrophe character has special meaning in XML, and must be substituted with "&apos;" in the value of an element. Thus "5'UTR" must be represented as "5&apos;UTR" in the XML file, i.e., `<INSDFeature_key>5&apos;UTR</INSDFeature_key>`.

### 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

**Optional qualifiers**

- allele
- gene
- gene_synonym
- map
- note
- operon
- standard_name

**Organism scope**

- prokaryotes

**Molecule scope**

- DNA

**References**

5.30. Feature Key - 35 signal

Definition: A conserved hexamer about 35 bp upstream of the start point of bacterial transcription units; consensus = TTGAC<sub>a</sub> or TGTTGACA

Optional qualifiers:
- allele
- gene
- gene_synonym
- note
- operon
- standard_name

Region scope: prokaryote

Molecule scope: DNA

References:
### SECTION 6: DESCRIPTION OF QUALIFIERS FOR NUCLEIC ACID SEQUENCES

This section contains the list of qualifiers to be used for features in nucleic acid sequences. The qualifiers are listed in alphabetic order.

Where a Value format of “none” is indicated in the description of a qualifier (e.g. germline), the INSDQualifier_value element must not be used.

PLEASE NOTE: Any qualifier value provided for a qualifier with a “free text” value format may require translation for National/Regional procedures.

<table>
<thead>
<tr>
<th>Qualifier</th>
<th>Definition</th>
<th>Value format</th>
<th>Example</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>allele</td>
<td>name of the allele for the given gene</td>
<td>free text (NOTE: this value may require translation for National/Regional procedures)</td>
<td><code>&lt;INSDQualifier_value&gt;adh1-1&lt;/INSDQualifier_value&gt;</code></td>
<td>all gene-related features (exon, CDS etc) for a given gene should share the same allele qualifier value; the allele qualifier value must, by definition, be different from the gene qualifier value; when used with the variation feature key, the allele qualifier value should be that of the variant.</td>
</tr>
<tr>
<td>anticodon</td>
<td>location of the anticodon of tRNA and the amino acid for which it codes</td>
<td>(pos:&lt;location&gt;,aa:&lt;amino_acid&gt;,seq:&lt;text&gt;) where &lt;location&gt; is the position of the anticodon and &lt;amino_acid&gt; is the three letter abbreviation for the amino acid encoded and &lt;text&gt; is the sequence of the anticodon</td>
<td><code>&lt;INSDQualifier_value&gt;(pos:34..36,aa:Phe,seq:aaa)&lt;/INSDQualifier_value&gt;</code></td>
<td></td>
</tr>
<tr>
<td>bound_moiety</td>
<td>name of the molecule/complex that may bind to the given feature</td>
<td>free text (NOTE: this value may require translation for National/Regional procedures)</td>
<td><code>&lt;INSDQualifier_value&gt;GAL4&lt;/INSDQualifier_value&gt;</code></td>
<td>multiple bound_moiety qualifiers are legal on “promoter” and “enhancer” features. A single bound_moiety qualifier is legal on the “misc_binding”, “oriT” and “protein_bind” features.</td>
</tr>
<tr>
<td>cell_line</td>
<td>cell line from which the sequence was obtained</td>
<td>free text (NOTE: this value may require translation for National/Regional procedures)</td>
<td><code>&lt;INSDQualifier_value&gt;MCF7&lt;/INSDQualifier_value&gt;</code></td>
<td></td>
</tr>
<tr>
<td>6.5. Qualifier</td>
<td>cell_type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>----------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Definition</td>
<td>cell type from which the sequence was obtained</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value format</td>
<td>free text</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Example</td>
<td><code>&lt;INSDQualifier_value&gt;leukocyte&lt;/INSDQualifier_value&gt;</code></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6.6. Qualifier</th>
<th>chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>chromosome (e.g. Chromosome number) from which the sequence was obtained</td>
</tr>
<tr>
<td>Value format</td>
<td>free text</td>
</tr>
<tr>
<td>Example</td>
<td><code>&lt;INSDQualifier_value&gt;1&lt;/INSDQualifier_value&gt;</code> <code>&lt;INSDQualifier_value&gt;x&lt;/INSDQualifier_value&gt;</code></td>
</tr>
</tbody>
</table>

Comment: a source feature must not contain more than one clone should be specified for a given source feature qualifier; where the sequence was obtained from multiple clones it may be further described in the feature table using the feature key misc_feature and a note qualifier to specify the multiple clones.

<table>
<thead>
<tr>
<th>6.7. Qualifier</th>
<th>clone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>clone from which the sequence was obtained</td>
</tr>
<tr>
<td>Value format</td>
<td>free text</td>
</tr>
<tr>
<td>Example</td>
<td><code>&lt;INSDQualifier_value&gt;lambda-hIL7.3&lt;/INSDQualifier_value&gt;</code></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6.8. Qualifier</th>
<th>clone_lib</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>clone library from which the sequence was obtained</td>
</tr>
<tr>
<td>Value format</td>
<td>free text</td>
</tr>
<tr>
<td>Example</td>
<td><code>&lt;INSDQualifier_value&gt;lambda-hIL7&lt;/INSDQualifier_value&gt;</code></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6.9. Qualifier</th>
<th>codon_start</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>indicates the offset at which the first complete codon of a coding feature can be found, relative to the first base of that feature.</td>
</tr>
<tr>
<td>Value format</td>
<td>1 or 2 or 3</td>
</tr>
<tr>
<td>Example</td>
<td><code>&lt;INSDQualifier_value&gt;2&lt;/INSDQualifier_value&gt;</code></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6.10. Qualifier</th>
<th>collected_by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>name of persons or institute who collected the specimen</td>
</tr>
<tr>
<td>Value format</td>
<td>free text</td>
</tr>
<tr>
<td>Example</td>
<td><code>&lt;INSDQualifier_value&gt;Dan Janzen&lt;/INSDQualifier_value&gt;</code></td>
</tr>
</tbody>
</table>
### 6.11. Qualifier: collection_date

**Definition:** date that the specimen was collected.

**Value format:** DD-Mmm-YYYY, MMM-DD, YYYY-MM, or YYYY

**Example:**
- `<INSDQualifier_value>`21-Oct-1952</INSDQualifier_value>
- `<INSDQualifier_value>`Oct-1952</INSDQualifier_value>
- `<INSDQualifier_value>`1952</INSDQualifier_value>`

**Comment:** Full date format DD-Mmm-YYYY is preferred; where day and/or month of collection is not known either MMM-YYYY or YYYY can be used. Three-letter month abbreviations can be one of the following: Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec.

### 6.12. Qualifier: compare

**Definition:** Reference details of an existing public INSD entry to which a comparison is made

**Value format:** [accession-number.sequence-version]

**Example:** `<INSDQualifier_value>AJ634337.1</INSDQualifier_value>`

**Comment:** This qualifier may be used on the following features: misc_difference, unsure, and variation. Multiple compare qualifiers with different contents are allowed within a single feature. This qualifier is not intended for large-scale annotation of variations, such as SNPs.

### 6.13. Qualifier: cultivar

**Definition:** cultivar (cultivated variety) of plant from which sequence was obtained

**Value format:** free text

**Example:** `<INSDQualifier_value>Nipponbare</INSDQualifier_value>`

**Comment:** 'cultivar' is applied solely to products of artificial selection; use the variety qualifier for natural, named plant and fungal varieties.

### 6.14. Qualifier: dev_stage

**Definition:** if the sequence was obtained from an organism in a specific developmental stage, it is specified with this qualifier

**Value format:** free text

**Example:** `<INSDQualifier_value>fourth instar larva</INSDQualifier_value>`
6.15. Qualifier: direction
Definition: direction of DNA replication
Value format: left, right, or both
where left indicates toward the 5’ end of the sequence (as presented) and right indicates toward the 3’ end
Example: <INSDQualifier_value>LEFT</INSDQualifier_value>
Comment: The values left, right, and both are permitted when the direction qualifier is used to annotate a rep_origin feature key. However, only left and right values are permitted when the direction qualifier is used to annotate an oriT feature key. The values are case-insensitive, i.e. both “RIGHT” and “right” are valid.

6.16. Qualifier: EC_number
Definition: Enzyme Commission number for enzyme product of sequence
Value format: free text
Example: <INSDQualifier_value>1.1.2.4</INSDQualifier_value> <INSDQualifier_value>1.1.2.-</INSDQualifier_value> <INSDQualifier_value>1.1.2.n</INSDQualifier_value>
Comment: valid values for EC numbers are defined in the list prepared by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) (published in Enzyme Nomenclature 1992, Academic Press, San Diego, or a more recent revision thereof). The format represents a string of four numbers separated by full stops; up to three numbers starting from the end of the string may be replaced by dash “-” to indicate uncertain assignment. Symbol “n” may be used in the last position instead of a number where the EC number is awaiting assignment. Please note that such incomplete EC numbers are not approved by NC-IUBMB.

6.17. Qualifier: ecotype
Definition: a population within a given species displaying genetically based, phenotypic traits that reflect adaptation to a local habitat
Value format: free text
Example: <INSDQualifier_value>Columbia</INSDQualifier_value>
Comment: an example of such a population is one that has adapted hairier than normal leaves as a response to an especially sunny habitat. 'Ecotype' is often applied to standard genetic stocks of Arabidopsis thaliana, but it can be applied to any sessile organism.
### 6.18. Qualifier: environmental_sample

**Definition:**
Identifies sequences derived by direct molecular isolation from a bulk environmental DNA sample (by PCR with or without subsequent cloning of the product, DGGE, or other anonymous methods) with no reliable identification of the source organism. Environmental samples include clinical samples, gut contents, and other sequences from anonymous organisms that may be associated with a particular host. They do not include endosymbionts that can be reliably recovered from a particular host, organisms from a readily identifiable but uncultured field sample (e.g., many cyanobacteria), or phytoplasmas that can be reliably recovered from diseased plants (even though these cannot be grown in axenic culture).

**Value format:**
None

**Comment:**
Used only with the source feature key; source feature keys containing the environmental_sample qualifier should also contain the isolation_source qualifier.

A sequence including the environmental_sample qualifier must not include the strain qualifier.

### 6.19. Qualifier: exception

**Definition:**
Indicates that the coding region cannot be translated using standard biological rules.

**Value format:**
One of the following controlled vocabulary phrases:
- RNA editing
- Rearrangement required for product
- Annotated by transcript or proteomic data

**Example:**
- `<INSDQualifier_value>RNA editing</INSDQualifier_value>`
- `<INSDQualifier_value>Rearrangement required for product</INSDQualifier_value>`

**Comment:**
Only to be used to describe biological mechanisms such as RNA editing; protein translation of a CDS with an exception qualifier will be different from the corresponding conceptual translation; must not be used where transl_except qualifier would be adequate, e.g. in case of stop codon completion use.

### 6.20. Qualifier: frequency

**Definition:**
Frequency of the occurrence of a feature.

**Value format:**
Free text representing the proportion of a population carrying the feature expressed as a fraction.

*(NOTE: this value may require translation for National/Regional procedures)*

**Example:**
- `<INSDQualifier_value>23/108</INSDQualifier_value>`
- `<INSDQualifier_value>1 in 12</INSDQualifier_value>`
- `<INSDQualifier_value>0.85</INSDQualifier_value>`

### 6.21. Qualifier: function

**Definition:**
Function attributed to a sequence.

**Value format:**
Free text.

*(NOTE: this value may require translation for National/Regional procedures)*

**Example:**
- `<INSDQualifier_value>essential for recognition of cofactor</INSDQualifier_value>`

**Comment:**
The function qualifier is used when the gene name and/or product name do not convey the function attributable to a sequence.
<table>
<thead>
<tr>
<th>6.22.</th>
<th>Qualifier</th>
<th>gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>symbol of the gene corresponding to a sequence region</td>
<td></td>
</tr>
<tr>
<td>Value format</td>
<td>free text</td>
<td></td>
</tr>
<tr>
<td>Example</td>
<td><code>&lt;INSDQualifier_value&gt;ilvE&lt;/INSDQualifier_value&gt;</code></td>
<td></td>
</tr>
<tr>
<td>Comment</td>
<td>Use gene qualifier to provide the gene symbol; use standard_name qualifier to provide the full gene name.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6.23.</th>
<th>Qualifier</th>
<th>gene_synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>synonymous, replaced, obsolete or former gene symbol</td>
<td></td>
</tr>
<tr>
<td>Value format</td>
<td>free text</td>
<td></td>
</tr>
<tr>
<td>Example</td>
<td><code>&lt;INSDQualifier_value&gt;Hox-3.3&lt;/INSDQualifier_value&gt;</code></td>
<td></td>
</tr>
<tr>
<td>Comment</td>
<td>used where it is helpful to indicate a gene symbol synonym; when the gene_synonym qualifier is used, a primary gene symbol must always be indicated in a gene qualifier</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6.24.</th>
<th>Qualifier</th>
<th>germline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>the sequence presented has not undergone somatic rearrangement as part of an adaptive immune response; it is the unrearranged sequence that was inherited from the parental germline</td>
<td></td>
</tr>
<tr>
<td>Value format</td>
<td>none</td>
<td></td>
</tr>
<tr>
<td>Comment</td>
<td>germline qualifier must not be used to indicate that the source of the sequence is a gamete or germ cell; germline and rearranged qualifiers must not be used in the same source feature; germline and rearranged qualifiers must only be used for molecules that can undergo somatic rearrangements as part of an adaptive immune response; these are the T-cell receptor (TCR) and immunoglobulin loci in the jawed vertebrates, and the unrelated variable lymphocyte receptor (VLR) locus in the jawless fish (lampreys and hagfish); germline and rearranged qualifiers should not be used outside of the Craniata (taxid=89593)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6.25.</th>
<th>Qualifier</th>
<th>haplogroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>name for a group of similar haplotypes that share some sequence variation. Haplogroups are often used to track migration of population groups.</td>
<td></td>
</tr>
<tr>
<td>Value format</td>
<td>free text</td>
<td></td>
</tr>
<tr>
<td>Example</td>
<td><code>&lt;INSDQualifier_value&gt;H*&lt;/INSDQualifier_value&gt;</code></td>
<td></td>
</tr>
<tr>
<td>6.26.</td>
<td>Qualifier</td>
<td>haplotype</td>
</tr>
<tr>
<td>-------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td><strong>Definition</strong></td>
<td>name for a specific set of alleles that are linked together on the same physical chromosome. In the absence of recombination, each haplotype is inherited as a unit, and may be used to track gene flow in populations.</td>
<td></td>
</tr>
<tr>
<td><strong>Value format</strong></td>
<td>free text</td>
<td></td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td><code>&lt;INSDQualifier_value&gt;Dw3 B5 Cw1 A1&lt;/INSDQualifier_value&gt;</code></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6.27.</th>
<th>Qualifier</th>
<th>host</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
<td>natural (as opposed to laboratory) host to the organism from which sequenced molecule was obtained</td>
<td></td>
</tr>
<tr>
<td><strong>Value format</strong></td>
<td>free text</td>
<td></td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td><code>&lt;INSDQualifier_value&gt;Homo sapiens&lt;/INSDQualifier_value&gt;</code></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6.28.</th>
<th>Qualifier</th>
<th>identified_by</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
<td>name of the expert who identified the specimen taxonomically</td>
<td></td>
</tr>
<tr>
<td><strong>Value format</strong></td>
<td>free text</td>
<td></td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td><code>&lt;INSDQualifier_value&gt;John Burns&lt;/INSDQualifier_value&gt;</code></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6.29.</th>
<th>Qualifier</th>
<th>isolate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
<td>individual isolate from which the sequence was obtained</td>
<td></td>
</tr>
<tr>
<td><strong>Value format</strong></td>
<td>free text</td>
<td></td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td><code>&lt;INSDQualifier_value&gt;Patient #152&lt;/INSDQualifier_value&gt;</code></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6.30.</th>
<th>Qualifier</th>
<th>isolation_source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
<td>describes the physical, environmental and/or local geographical source of the biological sample from which the sequence was derived</td>
<td></td>
</tr>
<tr>
<td><strong>Value format</strong></td>
<td>free text</td>
<td></td>
</tr>
<tr>
<td><strong>Examples</strong></td>
<td><code>&lt;INSDQualifier_value&gt;rumen isolates from standard Pelleted ration-fed steer #67&lt;/INSDQualifier_value&gt;</code></td>
<td></td>
</tr>
<tr>
<td><strong>Comment</strong></td>
<td>used only with the source feature key; source feature keys containing an environmental_sample qualifier should also contain an isolation_source qualifier</td>
<td></td>
</tr>
</tbody>
</table>
**6.31. Qualifier**  
**lab_host**  
**Definition**  
scientific name of the laboratory host used to propagate the source organism from which the sequenced molecule was obtained  
**Value format**  
free text  
*(NOTE: this value may require translation for National/Regional procedures)*  
**Example**  
<INSDQualifier_value>Gallus gallus</INSDQualifier_value>  
<INSDQualifier_value>Gallus gallus embryo</INSDQualifier_value>  
<INSDQualifier_value>Escherichia coli strain DH5 alpha</INSDQualifier_value>  
<INSDQualifier_value>Homo sapiens HeLa cells</INSDQualifier_value>  
**Comment**  
the full binomial scientific name of the host organism should be used when known; extra conditional information relating to the host may also be included

**6.32. Qualifier**  
**lat_lon**  
**Definition**  
geographical coordinates of the location where the specimen was collected  
**Value format**  
free text - degrees latitude and longitude in format "d[dd.dddd] N|S d[dd.dddd] W|E"  
*(NOTE: this value may require translation for National/Regional procedures)*  
**Example**  
<INSDQualifier_value>47.94 N 28.12 W</INSDQualifier_value>  
<INSDQualifier_value>45.0123 S 4.1234 E</INSDQualifier_value>  

**6.33. Qualifier**  
**macronuclear**  
**Definition**  
if the sequence shown is DNA and from an organism which undergoes chromosomal differentiation between macronuclear and micronuclear stages, this qualifier is used to denote that the sequence is from macronuclear DNA  
**Value format**  
none

**6.34. Qualifier**  
**map**  
**Definition**  
genomic map position of feature  
**Value format**  
free text  
*(NOTE: this value may require translation for National/Regional procedures)*  
**Example**  
<INSDQualifier_value>8q12-13</INSDQualifier_value>

**6.35. Qualifier**  
**mating_type**  
**Definition**  
mating type of the organism from which the sequence was obtained; mating type is used for prokaryotes, and for eukaryotes that undergo meiosis without sexually dimorphic gametes  
**Value format**  
free text  
*(NOTE: this value may require translation for National/Regional procedures)*  
**Examples**  
<INSDQualifier_value>MAT-1</INSDQualifier_value>  
<INSDQualifier_value>plus</INSDQualifier_value>  
<INSDQualifier_value>-</INSDQualifier_value>  
<INSDQualifier_value>odd</INSDQualifier_value>  
<INSDQualifier_value>even</INSDQualifier_value>  
**Comment**  
mating_type qualifier values male and female are valid in the prokaryotes, but not in the eukaryotes; for more information, see the entry for the sex qualifier.
### 6.36. Qualifier: mobile_element_type

**Definition:**
Type and name or identifier of the mobile element which is described by the parent feature.

**Value format:**
\(<\text{mobile_element_type}>[:<\text{mobile_element_name}>]\)
where `<mobile_element_type>` is one of the following:
- transposon
- retrotransposon
- integron
- insertion sequence
- non-LTR retrotransposon
- SINE
- MITE
- LINE
- other

**Example:**
\(<\text{INSDQualifier_value}>\text{transposon}:\text{Tnp9}</\text{INSDQualifier_value}>\)

**Comment:**
`mobile_element_type` is legal on `mobile_element` feature key only. Mobile element should be used to represent both elements which are currently mobile, and those which were mobile in the past. Value "other" for `<mobile_element_type>` requires a `<mobile_element_name>`.

### 6.37. Qualifier: mod_base

**Definition:**
Abbreviation for a modified nucleotide base.

**Value format:**
Modified base abbreviation chosen from this Annex, refer to Section 2.

**Example:**
\(<\text{INSDQualifier_value}>\text{m5c}</\text{INSDQualifier_value}>\)
\(<\text{INSDQualifier_value}>\text{OTHER}</\text{INSDQualifier_value}>\)

**Comment:**
Specific modified nucleotides not found in Section 2 of this Annex are annotated by entering OTHER as the value for the `mod_base` qualifier and including a note qualifier with the full name of the modified base as its value.
### mol_type

**Definition:** molecule type of sequence

**Value format:** One chosen from the following:
- genomic DNA
- genomic RNA
- mRNA
- tRNA
- rRNA
- other RNA
- other DNA
- transcribed RNA
- viral cRNA
- unassigned DNA
- unassigned RNA

**Example:**
- `<INSDQualifier_value>genomic DNA</INSDQualifier_value>`
- `<INSDQualifier_value>other RNA</INSDQualifier_value>`

**Comment:** mol_type qualifier is mandatory on the source feature key; the value “genomic DNA” does not imply that the molecule is nuclear (e.g. organelle and plasmid DNA must be described using “genomic DNA”); ribosomal RNA genes must be described using “genomic DNA”; “rRNA” must only be used if the ribosomal RNA molecule itself has been sequenced; values “other RNA” and “other DNA” must be applied to synthetic molecules, values “unassigned DNA”, “unassigned RNA” must be applied where in vivo molecule is unknown.

### ncRNA_class

**Definition:** a structured description of the classification of the non-coding RNA described by the ncRNA parent key

**Value format:** `TYPE` where `TYPE` is one of the following controlled vocabulary terms or phrases:
- antisense_RNA
- autocatalytically_spliced_intron
- ribozyme
- hammerhead_ribozyme
- lncRNA
- RNase_P_RNA
- RNase_MRP_RNA
- telomerase_RNA
- guide_RNA
- rasiRNA
- scRNA
- sRNA
- mIRNA
- piRNA
- snoRNA
- snRNA
- SRP_RNA
- vault_RNA
- Y_RNA
- other

**Example:**
- `<INSDQualifier_value>autocatalytically_spliced_intron</INSDQualifier_value>`
- `<INSDQualifier_value>sRNA</INSDQualifier_value>`
- `<INSDQualifier_value>scRNA</INSDQualifier_value>`
- `<INSDQualifier_value>other</INSDQualifier_value>`

**Comment:** specific ncRNA types not yet in the ncRNA_class controlled vocabulary must be annotated by entering “other” as the ncRNA_class qualifier value, and providing a brief explanation of novel ncRNA_class in a note qualifier
### Qualifier note

**Definition** any comment or additional information

**Value format** free text  
*NOTE: this value may require translation for National/Regional procedures*

**Example** <INSDQualifier_value>A comment about the feature</INSDQualifier_value>

### 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** free text (with no whitespace characters)  
*NOTE: this value may require translation for National/Regional procedures*

**Example** <INSDQualifier_value>4</INSDQualifier_value>  
<INSDQualifier_value>6B</INSDQualifier_value>

**Comment** text limited to integers, letters or combination of integers and/or letters represented as a data value that contains no whitespace characters; any additional terms should be included in a standard_name qualifier. Example: a number qualifier with a value of 2A and a standard_name qualifier with a value of “long”

### Qualifier operon

**Definition** name of the group of contiguous genes transcribed into a single transcript to which that feature belongs

**Value format** free text  
*NOTE: this value may require translation for National/Regional procedures*

**Example** <INSDQualifier_value>lac</INSDQualifier_value>

**Comment** valid only on Prokaryota-specific features
<table>
<thead>
<tr>
<th>Section</th>
<th>Qualifier</th>
<th>Definition</th>
<th>Value Format</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.43.</td>
<td>organelle</td>
<td>type of membrane-bound intracellular structure from which the sequence was obtained</td>
<td>One of the following controlled vocabulary terms and phrases: chromatophore hydrogenosome mitochondrion nucleomorph plastid mitochondrion:kinetoplast plastid:chloroplast plastid:apicoplast plastid:chromoplast plastid:cyanelle plastid:leucoplast plastid:proplastid</td>
<td>&lt;INSDQualifier_value&gt;chromatophore&lt;/INSDQualifier_value&gt; &lt;INSDQualifier_value&gt;hydrogenosome&lt;/INSDQualifier_value&gt; &lt;INSDQualifier_value&gt;mitochondrion&lt;/INSDQualifier_value&gt; &lt;INSDQualifier_value&gt;nucleomorph&lt;/INSDQualifier_value&gt; &lt;INSDQualifier_value&gt;plastid&lt;/INSDQualifier_value&gt; &lt;INSDQualifier_value&gt;mitochondrion:kinetoplast&lt;/INSDQualifier_value&gt; &lt;INSDQualifier_value&gt;plastid:chloroplast&lt;/INSDQualifier_value&gt; &lt;INSDQualifier_value&gt;plastid:apicoplast&lt;/INSDQualifier_value&gt; &lt;INSDQualifier_value&gt;plastid:chromoplast&lt;/INSDQualifier_value&gt; &lt;INSDQualifier_value&gt;plastid:cyanelle&lt;/INSDQualifier_value&gt; &lt;INSDQualifier_value&gt;plastid:leucoplast&lt;/INSDQualifier_value&gt; &lt;INSDQualifier_value&gt;plastid:proplastid&lt;/INSDQualifier_value&gt;</td>
</tr>
<tr>
<td>6.44.</td>
<td>organism</td>
<td>scientific name of the organism that provided the sequenced genetic material, if known, or the available taxonomic information if the organism is unclassified; or an indication that the sequence is a synthetic construct</td>
<td>free text</td>
<td>&lt;INSDQualifier_value&gt;Homo sapiens&lt;/INSDQualifier_value&gt;</td>
</tr>
</tbody>
</table>
6.45. **Qualifier**  
**Definition**  
PCR primers that were used to amplify the sequence. A single PCR_primers qualifier should contain all the primers used for a single PCR reaction. If multiple forward or reverse primers are present in a single PCR reaction, multiple sets of fwd_name/fwd_seq or rev_name/rev_seq values will be present.

**Value format**  

**Example**  
<INSDQualifier_value>fwd_name: CO1P1, fwd_seq: ttgatttttttgacggaagg, rev_name: CO1R4, rev_seq: cwwytardcctarraartttgccg</INSDQualifier_value>

**Comment**  
fwd_seq and rev_seq are both mandatory; fwd_name and rev_name are both optional. Both sequences must be presented in 5'→3' order. The sequences, which must be enclosed within angle brackets < >, in XML, the angle brackets < and > must be substituted with &lt; and &gt; since they are reserved characters in XML.

6.46. **Qualifier**  
**Definition**  
phenotype conferred by the feature, where phenotype is defined as a physical, biochemical or behavioural characteristic or set of characteristics

**Value format**  
free text  
(NOTE: this value may require translation for National/Regional procedures)

**Example**  
<INSDQualifier_value>erythromycin resistance</INSDQualifier_value>

6.47. **Qualifier**  
**Definition**  
name of naturally occurring plasmid from which the sequence was obtained, where plasmid is defined as an independently replicating genetic unit that cannot be described by chromosome or segment qualifiers

**Value format**  
free text  
(NOTE: this value may require translation for National/Regional procedures)

**Example**  
<INSDQualifier_value>pC589</INSDQualifier_value>

6.48. **Qualifier**  
**Definition**  
name of subpopulation or phenotype of the sample from which the sequence was derived

**Value format**  
free text  
(NOTE: this value may require translation for National/Regional procedures)

**Example**  
<INSDQualifier_value>pop1</INSDQualifier_value>  
<INSDQualifier_value>Bear Paw</INSDQualifier_value>
## Qualifier: product

**Definition**: name of the product associated with the feature, e.g. the mRNA of an mRNA feature, the polypeptide of a CDS, the mature peptide of a mat_peptide, etc.

**Value format**: free text

**Example**: `<INSDQualifier_value>trypsinogen</INSDQualifier_value>` (when qualifier appears in CDS feature)

`<INSDQualifier_value>trypsin</INSDQualifier_value>` (when qualifier appears in mat_peptide feature)

`<INSDQualifier_value>XYZ neural-specific transcript</INSDQualifier_value>` (when qualifier appears in mRNA feature)

## Qualifier: protein_id

**Definition**: protein sequence identification number, an integer used in a sequence listing to designate the protein sequence encoded by the coding sequence identified in the corresponding CDS feature key and translation qualifier.

**Value format**: an integer greater than zero

**Example**: `<INSDQualifier_value>89</INSDQualifier_value>`

## Qualifier: proviral

**Definition**: this qualifier is used to flag sequence obtained from a virus or phage that is integrated into the genome of another organism.

**Value format**: none

## Qualifier: pseudo

**Definition**: indicates that this feature is a non-functional version of the element named by the feature key.

**Value format**: none

**Comment**: The qualifier pseudo should be used to describe non-functional genes that are not formally described as pseudogenes, e.g. CDS has no translation due to other reasons than pseudogenization events. Other reasons may include sequencing or assembly errors. In order to annotate pseudogenes the qualifier pseudogene must be used, indicating the TYPE of pseudogene.
### 6.53. Qualifier

| Definition | pseudogene |

indicates that this feature is a pseudogene of the element named by the feature key.

| Value format | TYPE |

where TYPE is one of the following controlled vocabulary terms or phrases:
- processed
- unprocessed
- unitary
- allelic
- unknown

| Example | 

**<INSDQualifier_value>processed</INSDQualifier_value>**
**<INSDQualifier_value>unprocessed</INSDQualifier_value>**
**<INSDQualifier_value>unitary</INSDQualifier_value>**
**<INSDQualifier_value>allelic</INSDQualifier_value>**
**<INSDQualifier_value>unknown</INSDQualifier_value>**

| Comment | Definitions of TYPE values:

- **processed** - the pseudogene has arisen by reverse transcription of a mRNA into cDNA, followed by reintegration into the genome. Therefore, it has lost any intron/exon structure, and it might have a pseudo-polyA-tail.
- **unprocessed** - the pseudogene has arisen from a copy of the parent gene by duplication followed by accumulation of random mutations. The changes, compared to their functional homolog, include insertions, deletions, premature stop codons, frameshifts and a higher proportion of non-synonymous versus synonymous substitutions.
- **unitary** - the pseudogene has no parent. It is the original gene, which is functional is some species but disrupted in some way (indels, mutation, recombination) in another species or strain.
- **allelic** - a (unitary) pseudogene that is stable in the population but importantly it has a functional alternative allele also in the population. i.e., one strain may have the gene, another strain may have the pseudogene. MHC haplotypes have allelic pseudogenes.
- **unknown** - the submitter does not know the method of pseudogenisation.

### 6.54. Qualifier

| Definition | rearranged |

the sequence presented in the entry has undergone somatic rearrangement as part of an adaptive immune response; it is not the unrearranged sequence that was inherited from the parental germline.

| Value format | none |

| Comment | The rearranged qualifier **must** not be used to annotate chromosome rearrangements that are not involved in an adaptive immune response; germline and rearranged qualifiers **must** only be used for molecules that can undergo somatic rearrangements as part of an adaptive immune response; these are the T-cell receptor (TCR) and immunoglobulin loci in the jawed vertebrates, and the unrelated variable lymphocyte receptor (VLR) locus in the jawless fish (lampreys and hagfish); germline and rearranged qualifiers should not be used outside of the Craniata (taxid=89593).
6.55. **Qualifier** recombination_class

**Definition**
A structured description of the classification of recombination hotspot regions within a sequence.

**Value format** TYPE
Where TYPE is one of the following controlled vocabulary terms or phrases:
- mitotic_recombination
- non_allelic_homologous_recombination
- chromosome_breakpoint

**Example**
`<INSDQualifier_value>meiotic recombination</INSDQualifier_value>`
`<INSDQualifier_value>chromosome_breakpoint</INSDQualifier_value>`

**Comment**
Specific recombination classes not yet in the recombination_class controlled vocabulary must be annotated by entering “other” as the recombination_class qualifier value and providing a brief explanation of the novel recombination_class in a note qualifier.

6.56. **Qualifier** regulatory_class

**Definition**
A structured description of the classification of transcriptional, translational, replicational and chromatin structure related regulatory elements in a sequence.

**Value format** TYPE
Where TYPE is one of the following controlled vocabulary terms or phrases:
- DNase_I_hypersensitive_site
- enhancer_blocking_element
- imprinting_control_region
- insulator
- locus_control_region
- matrix_attachment_region
- minus_35_signal
- minus_10_signal
- recoding_stimulatory_region
- replication_regulatory_region
- response_element
- polyA_signal_sequence
- ribosome_binding_site
- riboswitch
- silencer
- TATA_box
- transcriptional_cis_regulatory_region
- other

**Example**
`<INSDQualifier_value>promoter</INSDQualifier_value>`
`<INSDQualifier_value>enhancer</INSDQualifier_value>`
`<INSDQualifier_value>ribosome_binding_site</INSDQualifier_value>`

**Comment**
Specific regulatory classes not yet in the regulatory_class controlled vocabulary must be annotated by entering “other” as the regulatory_class qualifier value and providing a brief explanation of the novel regulatory_class in a note qualifier.
6.57. Qualifier replace

Definition indicates that the sequence identified in a feature's location is replaced by the sequence shown in the qualifier's value; if no sequence (i.e., no value) is contained within the qualifier, this indicates a deletion

Value format free text

Example <INSDQualifier_value>a</INSDQualifier_value> &lt;INSDQualifier_value></INSDQualifier_value&gt; - for a deletion

6.58. Qualifier ribosomal_slippage

Definition during protein translation, certain sequences can program ribosomes to change to an alternative reading frame by a mechanism known as ribosomal slippage

Value format none

Comment a join operator, e.g.: [join(486..1784,1787..4810)] must be used in the CDS feature location to indicate the location of ribosomal_slippage

6.59. Qualifier rpt_family

Definition type of repeated sequence; "Alu" or "Kpn", for example

Value format free text

Example <INSDQualifier_value>Alu</INSDQualifier_value>

6.60. Qualifier rpt_type

Definition organization structure and distribution of repeated sequence

Value format One of the following controlled vocabulary terms or phrases:
tandem
direct
inverted
flanking
nested
dispersed

Example <INSDQualifier_value>INVERTED</INSDQualifier_value> &lt;INSDQualifier_value>long_terminal_repeat</INSDQualifier_value>

Comment the values are case insensitive; i.e. both "INVERTED" and "inverted" are valid; Definitions of the values:
tandem - a repeat that exists adjacent to another in the same orientation;
direct - a repeat that exists not always adjacent but is in the same orientation;
inverted - a repeat which occurs in part or in total normally in the reverse orientation to one another on the same molecule;
flanking - a repeat lying outside the sequence for which it has functional significance (e.g. transposon insertion target sites);

nested - a repeat that is disrupted by the insertion of another element;

dispersed - a repeat that is found dispersed throughout the genome;

terminal - a repeat at the ends of and within the sequence for which it has functional significance (e.g. transposon LTRs);

long terminal repeat - a sequence directly repeated at both ends of a defined sequence, of the sort typically found in retroviral LTRs;

non LTR retrotransposon polymeric tract - a polymeric tract, such as poly(dA), within a non LTR retrotransposon;

centromeric repeat - a repeat region found within the modular centromere;

telomeric repeat - a repeat region found within the telomere;

x element combinatorial repeat - a repeat region located between the X element and the telomere or adjacent Y' elements;

y prime element - a repeat region located adjacent to telomeric repeats or X element combinatorial repeats, either as a single copy or tandem repeat of two to four copies;

other - a repeat exhibiting important attributes that cannot be described by other values.

6.61. Qualifier rpt_unit_range

Definition location [range] of a repeating unit expressed as a range

Value format <base_range> - where <base_range> is the first and last base (separated by two dots) of a repeating unit

Example <INSDQualifier_value>202..245</INSDQualifier_value>

Comment used to indicate the base range of the sequence that constitutes a repeating unit within the region specified by the feature keys oriT and repeat_region.

6.62. Qualifier rpt_unit_seq

Definition identity of a repeat sequence

Value format free text

(Note: this value may require translation for National/Regional procedures)

Example <INSDQualifier_value>aagggc</INSDQualifier_value>

Example <INSDQualifier_value>ag(5)tg(8)</INSDQualifier_value>

Example <INSDQualifier_value>(AAAGA)6(AAAA)1(AAAGA)12</INSDQualifier_value>

Comment used to indicate the literal sequence that constitutes a repeating unit within the region specified by the feature keys oriT and repeat_region.
<table>
<thead>
<tr>
<th>6.63.</th>
<th>Qualifier</th>
<th>satellite</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Definition</td>
<td>identifier for a satellite DNA marker, compose of many tandem repeats (identical or related) of a short basic repeated unit</td>
</tr>
<tr>
<td></td>
<td>Value format</td>
<td><code>&lt;satellite_type&gt;[:&lt;class&gt;][&lt;identifier&gt;]</code> - where <code>&lt;satellite_type&gt;</code> is one of the following: satellite; microsatellite; minisatellite</td>
</tr>
<tr>
<td></td>
<td>Example</td>
<td><code>&lt;INSDQualifier_value&gt;satellite: S1a&lt;/INSDQualifier_value&gt;</code>&lt;br&gt;<code>&lt;INSDQualifier_value&gt;satellite: alpha&lt;/INSDQualifier_value&gt;</code>&lt;br&gt;<code>&lt;INSDQualifier_value&gt;satellite: gamma III&lt;/INSDQualifier_value&gt;</code>&lt;br&gt;<code>&lt;INSDQualifier_value&gt;microsatellite: DC130&lt;/INSDQualifier_value&gt;</code></td>
</tr>
<tr>
<td></td>
<td>Comment</td>
<td>many satellites have base composition or other properties that differ from those of the rest of the genome that allows them to be identified.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6.64.</th>
<th>Qualifier</th>
<th>segment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Definition</td>
<td>name of viral or phage segment sequenced</td>
</tr>
<tr>
<td></td>
<td>Value format</td>
<td>free text</td>
</tr>
<tr>
<td></td>
<td>Example</td>
<td><code>&lt;INSDQualifier_value&gt;6&lt;/INSDQualifier_value&gt;</code></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6.65.</th>
<th>Qualifier</th>
<th>serotype</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Definition</td>
<td>serological variety of a species characterized by its antigenic properties</td>
</tr>
<tr>
<td></td>
<td>Value format</td>
<td>free text</td>
</tr>
<tr>
<td></td>
<td>Example</td>
<td><code>&lt;INSDQualifier_value&gt;B1&lt;/INSDQualifier_value&gt;</code></td>
</tr>
<tr>
<td></td>
<td>Comment</td>
<td>used only with the source feature key; the Bacteriological Code recommends the use of the term 'serovar' instead of 'serotype' for the prokaryotes; see the International Code of Nomenclature of Bacteria (1990 Revision) Appendix 10.B &quot;Infraspecific Terms&quot;.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6.66.</th>
<th>Qualifier</th>
<th>serovar</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Definition</td>
<td>serological variety of a species (usually a prokaryote) characterized by its antigenic properties</td>
</tr>
<tr>
<td></td>
<td>Value format</td>
<td>free text</td>
</tr>
<tr>
<td></td>
<td>Example</td>
<td><code>&lt;INSDQualifier_value&gt;0157:H7&lt;/INSDQualifier_value&gt;</code></td>
</tr>
<tr>
<td></td>
<td>Comment</td>
<td>used only with the source feature key; the Bacteriological Code recommends the use of the term 'serovar' instead of 'serotype' for prokaryotes; see the International Code of Nomenclature of Bacteria (1990 Revision) Appendix 10.B &quot;Infraspecific Terms&quot;.</td>
</tr>
</tbody>
</table>
### Qualifier: sex

**Definition:** Sex of the organism from which the sequence was obtained; sex is used for eukaryotic organisms that undergo meiosis and have sexually dimorphic gametes.

**Value format:** Free text

**Examples:**
- `<INSQQualifier_value>female</INSQQualifier_value>`
- `<INSQQualifier_value>male</INSQQualifier_value>`
- `<INSQQualifier_value>hermaphrodite</INSQQualifier_value>`
- `<INSQQualifier_value>unisexual</INSQQualifier_value>`
- `<INSQQualifier_value>bisexual</INSQQualifier_value>`
- `<INSQQualifier_value>monoeccious</INSQQualifier_value>`
- `<INSQQualifier_value>dioecious</INSQQualifier_value>`

**Comment:**
The sex qualifier should be used (instead of mating_type qualifier) in the Metazoa, Embryophyta, Rhodophyta & Phaeophyceae; mating_type qualifier should be used (instead of sex qualifier) in the Bacteria, Archaea & Fungi; neither sex nor mating_type qualifiers should be used in the viruses; outside of the taxa listed above, mating_type qualifier should be used unless the value of the qualifier is taken from the vocabulary given in the examples above.

### Qualifier: standard_name

**Definition:** Accepted standard name for this feature.

**Value format:** Free text

**Example:** `<INSQQualifier_value>dotted</INSQQualifier_value>`

**Comment:** Use standard_name qualifier to give full gene name, but use gene qualifier to give gene symbol (in the above example gene qualifier value is Dt).

### Qualifier: strain

**Definition:** Strain from which sequence was obtained.

**Value format:** Free text

**Example:** `<INSQQualifier_value>BALB/c</INSQQualifier_value>`

**Comment:** Feature entries including a strain qualifier must not include the environmental_sample qualifier.

### Qualifier: sub_clone

**Definition:** Sub-clone from which sequence was obtained.

**Value format:** Free text

**Example:** `<INSQQualifier_value>lambda-hIL7.20g</INSQQualifier_value>`

**Comment:** A source feature must not contain more than one sub_clone qualifier; to indicate that the sequence was obtained from multiple sub-clones, multiple source features should be given. Sources may be further described using the feature key "misc-feature" and the qualifier "note."
<table>
<thead>
<tr>
<th>Qualifier</th>
<th>Definition</th>
<th>Value format</th>
<th>Example</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>sub_species</td>
<td>name of sub-species of organism from which sequence was obtained</td>
<td>free text</td>
<td><code>&lt;INSDQualifier_value&gt;lactis&lt;/INSDQualifier_value&gt;</code></td>
<td>(NOTE: this value may require translation for National/Regional procedures)</td>
</tr>
<tr>
<td>sub_strain</td>
<td>name or identifier of a genetically or otherwise modified strain from which sequence was obtained, derived from a parental strain (which should be annotated in the strain qualifier). sub_strain from which sequence was obtained</td>
<td>free text</td>
<td><code>&lt;INSDQualifier_value&gt;abis&lt;/INSDQualifier_value&gt;</code></td>
<td>(NOTE: this value may require translation for National/Regional procedures)</td>
</tr>
<tr>
<td>tag_peptide</td>
<td>base location encoding the polypeptide for proteolysis tag of tmRNA and its termination codon</td>
<td><code>&lt;base_range&gt;</code> - where <code>&lt;base_range&gt;</code> provides the first and last base (separated by two dots) of the location for the proteolysis tag</td>
<td><code>&lt;INSDQualifier_value&gt;90..122&lt;/INSDQualifier_value&gt;</code></td>
<td>it is recommended that the amino acid sequence corresponding to the tag_peptide be annotated by describing a 5’ partial CDS feature; e.g. CDS with a location of <code>&lt;90..122</code></td>
</tr>
<tr>
<td>tissue_lib</td>
<td>tissue library from which sequence was obtained</td>
<td>free text</td>
<td><code>&lt;INSDQualifier_value&gt;tissue library 772&lt;/INSDQualifier_value&gt;</code></td>
<td>(NOTE: this value may require translation for National/Regional procedures)</td>
</tr>
<tr>
<td>tissue_type</td>
<td>tissue type from which the sequence was obtained</td>
<td>free text</td>
<td><code>&lt;INSDQualifier_value&gt;liver&lt;/INSDQualifier_value&gt;</code></td>
<td>(NOTE: this value may require translation for National/Regional procedures)</td>
</tr>
</tbody>
</table>
6.76. Qualifier transl_except

**Definition**
translational exception: single codon the translation of which does not conform to genetic code defined by organism or transl_table.

**Value format**
(pos:location,aa:<amino_acid>) where <amino_acid> is the three letter abbreviation for the amino acid coded by the codon at the base_range position

**Example**
<INSDQualifier_value>(pos:213..215,aa:Trp) </INSDQualifier_value>
<INSDQualifier_value>(pos:462..464,aa:OTHER) </INSDQualifier_value>
<INSDQualifier_value>(pos:2017,aa:TERM) </INSDQualifier_value>
<INSDQualifier_value>(pos:2000..2001,aa:TERM) </INSDQualifier_value>
<INSDQualifier_value>(pos:X22222:15..17,aa:Ala) </INSDQualifier_value>

**Comment**
if the amino acid is not one of the specific amino acids listed in Section 3 of this Annex, use OTHER as <amino_acid> and provide the name of the unusual amino acid in a note qualifier; for modified amino-acid selenocysteine use three letter abbreviation "sec" (one letter symbol "U" in amino-acid sequence) for <amino_acid>; for modified amino-acid pyrrolysine use three letter abbreviation "pyl" (one letter symbol "O" in amino-acid sequence) for <amino_acid>; for partial termination codons where TAA stop codon is completed by the addition of 3’ A residues to the mRNA either a single base_position or a base_range is used for the location, see the third and fourth examples above, in conjunction with a note qualifier indicating 'stop codon completed by the addition of 3’ A residues to the RNA'.

6.77. Qualifier transl_table

**Definition**
definition of genetic code table used if other than universal or standard genetic code table. Tables used are described in this Annex

**Value format**
<integer> where <integer> is the number assigned to the genetic code table

**Example**
<INSDQualifier_value>3</INSDQualifier_value> - example where the yeast mitochondrial code is to be used

**Comment**
if the transl_table qualifier is not used to further annotate a CDS feature key, then the CDS is translated using the Standard Code (i.e. Universal Genetic Code). Genetic code exceptions outside the range of specified tables are reported in transl_except qualifiers.

6.78. Qualifier trans_splicing

**Definition**
indicates that exons from two RNA molecules are ligated in intermolecular reaction to form mature RNA

**Value format**
none

**Comment**
should be used on features such as CDS, mRNA and other features that are produced as a result of a trans-splicing event. This qualifier <splicing_event> must be used only when the splice event is indicated in the "join" operator, e.g. join(complement(69611..69724),139856..140087) in the feature location
6.79. Qualifier  translation

Definition  one-letter abbreviated amino acid sequence derived from either the standard (or universal) genetic code or the table as specified in a transl_table qualifier and as determined by an exception in the transl_except qualifier.

Value format  contiguous string of one-letter amino acid abbreviations from Section 3 of this Annex, "X" is to be used for AA exceptions.

Example  <INSDQualifier_value>MASTFPPWYRCSTPSLKLIMCTW</INSDQualifier_value>

Comment  to be used with CDS feature only; must be accompanied by protein_id qualifier when the translation product contains four or more specifically defined amino acids; see transl_table for definition and location of genetic code Tables; only one of the qualifiers translation, pseudo and pseudogene are permitted to further annotate a CDS feature.

6.80. Qualifier  variety

Definition  variety (= varietas, a formal Linnaean rank) of organism from which sequence was derived.

Value format  free text

Example  <INSDQualifier_value>insularis</INSDQualifier_value>

Comment  use the cultivar qualifier for cultivated plant varieties, i.e., products of artificial selection; varieties other than plant and fungal varietas should be annotated via a note qualifier, e.g. with the value <INSDQualifier_value>breed:Cukorova</INSDQualifier_value>
SECTION 7: FEATURE KEYS FOR AMINO ACID SEQUENCES

This section contains the list of allowed feature keys to be used for amino acid sequences. The feature keys are listed in alphabetic order.

7.1. Feature Key

**ACT_SITE**

**Definition**
Amino acid(s) involved in the activity of an enzyme

**Optional qualifiers**
NOTE

**Comment**
Each amino acid **residue** of the active site **must** be annotated separately with the ACT_SITE feature key. The corresponding amino acid residue number **must** be provided as the location descriptor in the feature location element.

7.2. Feature Key

**BINDING**

**Definition**
Binding site for any chemical group (co-enzyme, prosthetic group, etc.). The chemical nature of the group is indicated in the NOTE qualifier

**Mandatory qualifiers**
NOTE

**Comment**
Examples of values for the “NOTE” qualifier: “Heme (covalent)” and “Chloride.” Where appropriate, the features keys CA_BIND, DNA_BIND, METAL, and NP_BIND should be used rather than BINDING.

7.3. Feature Key

**CA_BIND**

**Definition**
Extent of a calcium-binding region

**Optional qualifiers**
NOTE

7.4. Feature Key

**CARBOHYD**

**Definition**
Glycosylation site

**Mandatory qualifiers**
NOTE

**Comment**
This key describes the occurrence of the attachment of a glycan (mono- or polysaccharide) to a residue of the protein. The type of linkage (C-, N- or O-linked) to the protein is indicated in the “NOTE” qualifier. If the nature of the reducing terminal sugar is known, its abbreviation is shown between parentheses. If three dots ‘...’ follow the abbreviation this indicates an extension of the carbohydrate chain. Conversely no dots means that a monosaccharide is linked. Examples of values used in the “NOTE” qualifier: **O-linked (GlcNAc...)**, **O-linked (GlcNAc); C-linked (Glc...)**, **C-linked (Man...)**, **partial O-linked (Gal...)**.

7.5. Feature Key

**CHAIN**

**Definition**
Extent of a polypeptide chain in the mature protein

**Optional qualifiers**
NOTE
### Feature Key: COILED

**Definition:** Extent of a coiled-coil region

**Optional qualifiers:** NOTE

### Feature Key: COMPBIAS

**Definition:** Extent of a compositionally biased region

**Optional qualifiers:** NOTE

### Feature Key: CONFLICT

**Definition:** Different sources report differing sequences

**Optional qualifiers:** NOTE

**Comment:** Examples of values for the "NOTE" qualifier: Missing; K -> Q; GSDSE -> RIRLR; V -> A.

### Feature Key: CROSSLINK

**Definition:** Post translationally formed amino acid bonds

**Mandatory qualifiers:** NOTE

**Comment:** Covalent linkages of various types formed between two proteins (interchain cross-links) or between two parts of the same protein (intrachain cross-links); except for cross-links formed by disulfide bonds, for which the "DISULFID" feature key is to be used. For an interchain cross-link, the location descriptor in the feature location element is the residue number of the amino acid cross-linked to the other protein. For an intrachain cross-link, the location descriptors in the feature location element are the residue numbers of the cross-linked amino acids in conjunction with the "join" location operator, e.g. "join(42,50)." The NOTE qualifier indicates the nature of the cross-link, at least specifying the name of the conjugate and the identity of the two amino acids involved. Examples of values for the "NOTE" qualifier: "Isoglutamyl cysteine thioester (Cys-Gln);" "beta-methylanthionine (Cys-Thr);" and "Glycyl lysine isopeptide (Lys-Gly) (interchain with G-cte in ubiquitin)"

### Feature Key: DISULFID

**Definition:** Disulfide bond

**Optional qualifiers:** NOTE

**Comment:** For an interchain disulfide bond, the location descriptor in the feature location element is the residue number of the cystein linked to the other protein. For an intrachain cross-link, the location descriptors in the feature location element are the residue numbers of the linked cysteines in conjunction with the "join" location operator, e.g. "join(42,50)." For interchain disulfide bonds, the NOTE qualifier indicates the nature of the cross-link, by identifying the other protein, for example, "Interchain (between A and B chains)"
### DNA BIND
- **Definition:** Extent of a DNA-binding region
- **Mandatory qualifiers:** NOTE
- **Comment:** The nature of the DNA-binding region is given in the NOTE qualifier. Examples of values for the “NOTE” qualifier: “Homeobox” and “Myb 2”

### DOMAIN
- **Definition:** Extent of a domain, which is defined as a specific combination of secondary structures organized into a characteristic three-dimensional structure or fold
- **Mandatory qualifiers:** NOTE
- **Comment:** The domain type is given in the NOTE qualifier. Where several copies of a domain are present, the domains are numbered. Examples of values for the “NOTE” qualifier: “Ras-GAP” and “Cadherin 1”

### HELIX
- **Definition:** Secondary structure: Helices, for example, Alpha-helix; 3(10) helix; or Pi-helix
- **Optional qualifiers:** NOTE
- **Comment:** This feature is used only for proteins whose tertiary structure is known. Only three types of secondary structure are specified: helices (key HELIX), beta-strands (key STRAND) and turns (key TURN). Residues not specified in one of these classes are in a 'loop' or 'random-coil' structure.

### INIT MET
- **Definition:** Initiator methionine
- **Optional qualifiers:** NOTE
- **Comment:** The location descriptor in the feature location element is “1”. This feature key indicates the N-terminal methionine is cleaved off. This feature is not used when the initiator methionine is not cleaved off.

### INTRAMEM
- **Definition:** Extent of a region located in a membrane without crossing it
- **Optional qualifiers:** NOTE

### LIPID
- **Definition:** Covalent binding of a lipid moiety
- **Mandatory qualifiers:** NOTE
- **Comment:** The chemical nature of the bound lipid moiety is given in the NOTE qualifier, indicating at least the name of the lipidated amino acid. Examples of values for the “NOTE” qualifier: “N-myristoyl glycine”; “GPI-anchor amidated serine” and “s-diacylglycerol cysteine.”
7.17. Feature Key METAL
   Definition Binding site for a metal ion.
   Mandatory qualifiers NOTE
   Comment The NOTE qualifier indicates the nature of the metal. Examples of values for the
   “NOTE” qualifier: “Iron [heme axial ligand]” and “Copper”.

7.18. Feature Key MOD_RES
   Definition Posttranslational modification of a residue
   Mandatory qualifiers NOTE
   Comment The chemical nature of the modified residue is given in the NOTE qualifier,
   indicating at least the name of the post-translationally modified amino acid. If
   the modified amino acid is listed in Annex section 4 of this Annex, the abbreviation
   may be used in place of the full name. Examples of values for the “NOTE”
   qualifier: “N-acetylalanine”; “3-Hyp”; and “Melys” or “N-6-methyllysine”

7.19. Feature Key MOTIF
   Definition Short (up to 20 amino acids) sequence motif of biological interest
   Optional qualifiers NOTE

7.20. Feature Key MUTAGEN
   Definition Site which has been experimentally altered by mutagenesis
   Optional qualifiers NOTE

7.21. Feature Key NON_STD
   Definition Non-standard amino acid
   Optional qualifiers NOTE
   Comment This key describes the occurrence of non-standard amino acids selenocysteine (U)
   and pyrrolysine (O) in the amino acid sequence.

7.22. Feature Key NON_TER
   Definition The residue at an extremity of the sequence is not the terminal residue
   Optional qualifiers NOTE
   Comment If applied to position 1, this means that the first position is not the N-terminus
   of the complete molecule. If applied to the last position, it means that this
   position is not the C-terminus of the complete molecule.
7.23. Feature Key | NP_BIND  
Definition |Extent of a nucleotide phosphate-binding region  
Mandatory qualifiers |NOTE  
Comment |The nature of the nucleotide phosphate is indicated in the NOTE qualifier. Examples of values for the “NOTE” qualifier: “ATP” and “FAD”.

7.24. Feature Key | PEPTIDE  
Definition |Extent of a released active peptide  
Optional qualifiers |NOTE

7.25. Feature Key | PROPEP  
Definition |Extent of a propeptide  
Optional qualifiers |NOTE

7.26. Feature Key | REGION  
Definition |Extent of a region of interest in the sequence  
Optional qualifiers |NOTE

7.27. Feature Key | REPEAT  
Definition |Extent of an internal sequence repetition  
Optional qualifiers |NOTE

7.28. Feature Key | SIGNAL  
Definition |Extent of a signal sequence (prepeptide)  
Optional qualifiers |NOTE

7.29. Feature Key | SITE  
Definition |Any interesting single amino-acid site on the sequence that is not defined by another feature key. It can also apply to an amino acid bond which is represented by the positions of the two flanking amino acids  
Mandatory qualifier |NOTE  
Comment |When SITE is used to annotate a modified amino acid the value for the qualifier “NOTE” must either be an abbreviation set forth in Section 4 of this Annex, [NAME] for the complete, unabbreviated name of the modified amino acid.
7.30. Feature Key SOURCE
Definition Identifies the source of the sequence; this key is mandatory; every sequence will have a single SOURCE feature spanning the entire sequence
Mandatory qualifiers MOL_TYPE ORGANISM
Optional qualifiers NOTE

7.31. Feature Key STRAND
Definition Secondary structure: Beta-strand; for example Hydrogen bonded beta-strand or residue in an isolated beta-bridge
Optional qualifiers NOTE
Comment This feature is used only for proteins whose tertiary structure is known. Only three types of secondary structure are specified: helices (key HELIX), beta-strands (key STRAND) and turns (key TURN). Residues not specified in one of these classes are in a 'loop' or 'random-coil' structure.

7.32. Feature Key TOPO_DOM
Definition Topological domain
Optional qualifiers NOTE

7.33. Feature Key TRANS_MEM
Definition Extent of a transmembrane region
Optional qualifiers NOTE

7.34. Feature Key TRANSIT
Definition Extent of a transit peptide (mitochondrion, chloroplast, thylakoid, cyanelle, peroxisome etc.)
Optional qualifiers NOTE

7.35. Feature Key TURN
Definition Secondary structure Turns, for example, H-bonded turn (3-turn, 4-turn or 5-turn)
Optional qualifiers NOTE
Comment This feature is used only for proteins whose tertiary structure is known. Only three types of secondary structure are specified: helices (key HELIX), beta-strands (key STRAND) and turns (key TURN). Residues not specified in one of these classes are in a 'loop' or 'random-coil' structure.
<table>
<thead>
<tr>
<th>Feature Key</th>
<th>Definition</th>
<th>Optional qualifiers</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNSURE</td>
<td>Uncertainties in the amino acid sequence</td>
<td>NOTE</td>
<td>Used to describe region(s) of an amino acid sequence for which the authors are unsure about the sequence presentation.</td>
</tr>
<tr>
<td>VARIANT</td>
<td>Authors report that sequence variants exist</td>
<td>NOTE</td>
<td></td>
</tr>
<tr>
<td>VAR_SEQ</td>
<td>Description of sequence variants produced by alternative splicing, alternative promoter usage, alternative initiation and ribosomal frameshifting</td>
<td>NOTE</td>
<td></td>
</tr>
<tr>
<td>ZN_FING</td>
<td>Extent of a zinc finger region</td>
<td>NOTE</td>
<td>The type of zinc finger is indicated in the NOTE qualifier. For example: &quot;GATA-type&quot; and &quot;NR C4-type&quot;</td>
</tr>
</tbody>
</table>
## SECTION 8: QUALIFIERS FOR AMINO ACID SEQUENCES

This section contains the list of allowed qualifiers to be used for amino acid sequences.

**PLEASE NOTE:** Any qualifier value provided for a qualifier with a "free text" value format may require translation for National/Regional procedures.

<table>
<thead>
<tr>
<th>8.1. Qualifier</th>
<th>MOL_TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>In vivo molecule type of sequence</td>
</tr>
<tr>
<td>Value format</td>
<td>protein</td>
</tr>
<tr>
<td>Example</td>
<td><code>&lt;INSDQualifier_value&gt;protein&lt;/INSDQualifier_value&gt;</code></td>
</tr>
<tr>
<td>Comment</td>
<td>The &quot;MOL_TYPE&quot; qualifier is mandatory on the SOURCE feature key.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>8.2. Qualifier</th>
<th>NOTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>Any comment or additional information</td>
</tr>
<tr>
<td>Value format</td>
<td>free text</td>
</tr>
<tr>
<td>Example</td>
<td><code>&lt;INSDQualifier_value&gt;Heme (covalent)&lt;/INSDQualifier_value&gt;</code></td>
</tr>
<tr>
<td>Comment</td>
<td>The &quot;NOTE&quot; qualifier is mandatory for the feature keys: BINDING; CARBOHYD; CROSSLINK; DISULFID; DNA_BIND; DOMAIN; LIPID; METAL; MOD_RES; NP_BIND and ZN_FING</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>8.3. Qualifier</th>
<th>ORGANISM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>Scientific name of the organism that provided the peptide</td>
</tr>
<tr>
<td>Value format</td>
<td>free text</td>
</tr>
<tr>
<td>Example</td>
<td><code>&lt;INSDQualifier_value&gt;Homo sapiens&lt;/INSDQualifier_value&gt;</code></td>
</tr>
<tr>
<td>Comment</td>
<td>The &quot;ORGANISM&quot; qualifier is mandatory for the SOURCE feature key.</td>
</tr>
</tbody>
</table>
SECTION 9: GENETIC CODE TABLES

Table 5 reproduces [genetic code tables] to be used for translating coding sequences. The value for the trans_table qualifier is the number assigned to the corresponding genetic code table. Where a CDS feature is described with a translation qualifier but not a transl_table qualifier, the 1 - Standard Code is used by default for translation. (Note: Genetic code tables 7, 8, 15, and 17 to 20 do not exist, therefore these numbers do not appear in Table 5.)

<table>
<thead>
<tr>
<th>Table 5: Genetic Code Tables</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 - Standard Code</strong></td>
</tr>
<tr>
<td>AAs = FFLSSSY**CC*WLLPPPHHQRRRIIMTTTNNKSSSRVVVAADDEEGGGG</td>
</tr>
<tr>
<td>Starts = ----M---------------------M---------------------M</td>
</tr>
<tr>
<td>Base1 = ttttttttttttttttccccccccccccccccccccccccccccccccccccccc</td>
</tr>
<tr>
<td>Base2 = ttttttttttttttttccccccccccccccccccccccccccccccccccccccc</td>
</tr>
<tr>
<td>Base3 = tcagtcatgcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag</td>
</tr>
<tr>
<td><strong>2 - Vertebrate Mitochondrial Code</strong></td>
</tr>
<tr>
<td>AAs = FFLSSSY**CCWLLLPPPPPHHQRRRIIMTTTNNKSSSRVVVAADDEEGGGG</td>
</tr>
<tr>
<td>Starts = ------------------------M------------------------M</td>
</tr>
<tr>
<td>Base1 = ttttttttttttttttccccccccccccccccccccccccccccccccccccccc</td>
</tr>
<tr>
<td>Base2 = ttttttttttttttttccccccccccccccccccccccccccccccccccccccc</td>
</tr>
<tr>
<td>Base3 = tcagtcatgcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag</td>
</tr>
<tr>
<td><strong>3 - Yeast Mitochondrial Code</strong></td>
</tr>
<tr>
<td>AAs = FFLSSSY**CCWTTTTTTPPPPHHQRRRIIMTTTNNKSSSRVVVAADDEEGGGG</td>
</tr>
<tr>
<td>Starts = ------------------------M------------------------M</td>
</tr>
<tr>
<td>Base1 = ttttttttttttttttccccccccccccccccccccccccccccccccccccccc</td>
</tr>
<tr>
<td>Base2 = ttttttttttttttttccccccccccccccccccccccccccccccccccccccc</td>
</tr>
<tr>
<td>Base3 = tcagtcatgcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag</td>
</tr>
<tr>
<td><strong>4 - Mold, Protozoan, Coelenterate Mitochondrial Code &amp; Mycoplasma/Spiroplasma Code</strong></td>
</tr>
<tr>
<td>AAs = FFLSSSY**CCWLLLPPPPPHHQRRRIIMTTTNNKSSSRVVVAADDEEGGGG</td>
</tr>
<tr>
<td>Starts = ------------------------M------------------------M</td>
</tr>
<tr>
<td>Base1 = ttttttttttttttttccccccccccccccccccccccccccccccccccccccc</td>
</tr>
<tr>
<td>Base2 = ttttttttttttttttccccccccccccccccccccccccccccccccccccccc</td>
</tr>
<tr>
<td>Base3 = tcagtcatgcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag</td>
</tr>
<tr>
<td><strong>5 - Invertebrate Mitochondrial Code</strong></td>
</tr>
<tr>
<td>AAs = FFLSSYY**CCWLLLPPPPPHHQRRRIIMTTTNNKSSSRVVVAADDEEGGGG</td>
</tr>
<tr>
<td>Starts = ------------------------M------------------------M</td>
</tr>
<tr>
<td>Base1 = ttttttttttttttttccccccccccccccccccccccccccccccccccccccc</td>
</tr>
<tr>
<td>Base2 = ttttttttttttttttccccccccccccccccccccccccccccccccccccccc</td>
</tr>
<tr>
<td>Base3 = tcagtcatgcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag</td>
</tr>
<tr>
<td><strong>6 - Ciliate, Dasycladacean and Hexamita Nuclear Code</strong></td>
</tr>
<tr>
<td>AAs = FFLSSYYQCC*WLLLPPPPPHHQRRRIIMTTTNNKSSSRVVVAADDEEGGGG</td>
</tr>
<tr>
<td>Starts = ------------------------M------------------------M</td>
</tr>
<tr>
<td>Base1 = ttttttttttttttttccccccccccccccccccccccccccccccccccccccc</td>
</tr>
<tr>
<td>Base2 = ttttttttttttttttccccccccccccccccccccccccccccccccccccccc</td>
</tr>
<tr>
<td>Base3 = tcagtcatgcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag</td>
</tr>
<tr>
<td><strong>9 - Echinoderm and Flatworm Mitochondrial Code</strong></td>
</tr>
<tr>
<td>AAs = FFLSSYY**CCWLLLPPPPPHHQRRRIIMTTTNNKSSSRVVVAADDEEGGGG</td>
</tr>
<tr>
<td>Starts = ------------------------M------------------------M</td>
</tr>
<tr>
<td>Base1 = ttttttttttttttttccccccccccccccccccccccccccccccccccccccc</td>
</tr>
<tr>
<td>Base2 = ttttttttttttttttccccccccccccccccccccccccccccccccccccccc</td>
</tr>
<tr>
<td>Base3 = tcagtcatgcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag</td>
</tr>
<tr>
<td>Code Type</td>
</tr>
<tr>
<td>------------------------</td>
</tr>
</tbody>
</table>
| **10 - Euplotid Nuclear Code** | FFLSSSSSY**CCWLLLPPPPHQQRRRIIMTTTTNNKSSRRVVVVAAADDEEGGGG | tttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
<table>
<thead>
<tr>
<th>Code</th>
<th>AAs</th>
<th>Starts</th>
<th>Base1</th>
<th>Base2</th>
<th>Base3</th>
</tr>
</thead>
</table>
| 22 - Scenedesmus obliquus Mitochondrial Code | FFLS S*SYY*LCC*WLLL LPPPPHHQQRRIIM TTTT TNNKKSSRRVV VV VAAAADDEEGGG | M                         | tttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
30 - Peritrich Nuclear

\[
\begin{align*}
\text{AAs} &= \text{FFLLSSSYEECC*WLLLAPPQHRQRIITTTTNNKSSRRVVVAAADDEEGGGG} \\
\text{Starts} &= \text{---------------------------} \\
\text{Base1} &= \text{ttttttttttttttttcccccccccccccccccccccc} \\
\text{Base2} &= \text{ttttccccaaaaggg} \\
\text{Base3} &= \text{tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag}
\end{align*}
\]

31 - Blastocrithidia Nuclear

\[
\begin{align*}
\text{AAs} &= \text{FFLLSSSYEECCWLLLLPPPHQRRRIITTTTNNKSSRRVVVAAADDEEGGGG} \\
\text{Starts} &= \text{---------------------------} \\
\text{Base1} &= \text{ttttttttttttttttcccccccccccccccccccccc} \\
\text{Base2} &= \text{ttttccccaaaaggg} \\
\text{Base3} &= \text{tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag}
\end{align*}
\]

[Annex II follows]
ANNEX II

DOCUMENT TYPE DEFINITION (DTD) FOR SEQUENCE LISTING

Version 1.1.1

Adopted Approved by the Committee on WIPO Standards (CWS)
at its reconvened fourth session on March 24, 2016 fifth session on June 2, 2017

<?xml version="1.0" encoding="UTF-8"?>
<!--Annex II of WIPO Standard ST.26, Document Type Definition (DTD) for Sequence Listing

This entity may be identified by the PUBLIC identifier:
*******************************************************************************************
* PUBLIC "//WIPO//DTD SEQUENCE LISTING 1.1//EN" "ST26SequenceListing_V1_1.dtd"
*******************************************************************************************
* PUBLIC DTD URL
* http://www.wipo.int/standards/DTD/ST26SequenceListing_V1_1.dtd

WIP Standard ST.26, version 1.0, Recommended Standard for the presentation of nucleotide and amino acid sequence listings using XML (eXtensible Markup Language), adopted by the Committee on WIPO Standards (CWS) at its reconvened fourth session on March 24, 2016 Revision of Annex II to WIPO Standard ST.26 was approved by the Committee on WIPO Standards (CWS) at its fifth session.

* CONTACTS
xml.standards@wipo.int

* NOTES
The sequence data part is a subset of the complete INSDC DTD V.1.5 that only covers the requirements of WIPO Standard ST.26.

* REVISION HISTORY
2017-06-02: Version 1.1 approved at the CWS/5 Changes:
Comments added to <INSDSeq_length>, <INSDSeq_division> and <INSDSeq_sequence> to clarify the reason of the differences between the INSDC DTD V.1.5 and ST26 Sequence Listing DTD V1.1.

2016-03-24: Version 1.0 adopted at the CWS/4Bis
2014-03-11: Final draft for adoption.
ST26SequenceListing
*******************************************************************************
* ROOT ELEMENT
*******************************************************************************

<!ELEMENT ST26SequenceListing ((ApplicantFileReference | (ApplicationIdentification, ApplicantFileReference?), EarliestPriorityApplicationIdentification?, (ApplicantName, ApplicantNameLatin?), (InventorName, InventorNameLatin?), InventionTitle+, SequenceTotalQuantity, SequenceData+)>  

/* The elements ApplicantName and InventorName are optional in this DTD to facilitate the conversion between various encoding schemes */

<!ATTLIST ST26SequenceListing
dtdVersion CDATA #REQUIRED
fileName CDATA #IMPLIED
softwareName CDATA #IMPLIED
softwareVersion CDATA #IMPLIED
productionDate CDATA #IMPLIED>

<!-- ApplicantFileReference
Applicant's or agent's file reference, mandatory if application identification not provided. -->

<!ELEMENT ApplicantFileReference (#PCDATA)>

<!-- ApplicationIdentification
Application identification for which the sequence listing is submitted, when available. -->

<!ELEMENT ApplicationIdentification (IPOfficeCode, ApplicationNumberText, FilingDate?)>

<!-- EarliestPriorityApplicationIdentification
Application identification of the earliest claimed priority, which contains IPOfficeCode, ApplicationNumberText and FilingDate elements. For details, please see ApplicationIdentification. -->

<!ELEMENT EarliestPriorityApplicationIdentification (IPOfficeCode, ApplicationNumberText, FilingDate?)>

<!-- ApplicantName
The name of the first mentioned applicant in characters set forth in paragraph 40 (a) of the ST.26 main body document. -->

<!languageCode: Appropriate language code from ISO 639-1 - Codes for the representation of names of languages - Part 1: Alpha-2 -->

<!ELEMENT ApplicantName (#PCDATA)>
<!ATTLIST ApplicantName
languageCode CDATA #REQUIRED >

<!-- ApplicantNameLatin
Where ApplicantName is typed in characters other than those as set forth in paragraph 40 (b), a translation or transliteration of the name of the first mentioned applicant must also be typed in characters as set forth in paragraph 40 (b) of the ST.26 main body document. -->

<!ELEMENT ApplicantNameLatin (#PCDATA)>

<!-- InventorName
Name of the first mentioned inventor typed in the characters as set forth in paragraph 40 (a). -->

<!languageCode: Appropriate language code from ISO 639-1 - Codes for the representation of names of languages - Part 1: Alpha-2 -->

<!ELEMENT InventorName (#PCDATA)>
<!ATTLIST InventorName
    languageCode CDATA #REQUIRED >

<!--InventorName
Where InventorName is typed in characters other than those as set forth in paragraph 40
(b), a translation or transliteration of the first mentioned inventor may also be typed in
characters as set forth in paragraph 40 (b).
-->
<!ELEMENT InventorNameLatin (#PCDATA)>

<!--InventionTitle
Title of the invention typed in the characters as set forth in paragraph 40 (a) in the
language of filing. A translation of the title of the invention into additional languages
may be typed in the characters as set forth in paragraph 40 (a) using additional
InventionTitle elements. Preferably two to seven words.
-->
<!--languageCode: Appropriate language code from ISO 639-1 - Codes
for the representation of names of languages - Part 1: Alpha-2
-->
<!ELEMENT InventionTitle (#PCDATA)>
<!ATTLIST InventionTitle
    languageCode CDATA #REQUIRED >

<!--SequenceTotalQuantity
Indicates the total number of sequences in the document.
Its purpose is to be quickly accessible for automatic processing.
-->
<!ELEMENT SequenceTotalQuantity (#PCDATA)>

<!--SequenceData
Data for individual Sequence.
For intentionally skipped sequences see the ST.26 main body document.
-->
<!ELEMENT SequenceData (INSDSeq)>
<!ATTLIST SequenceData
    sequenceIDNumber CDATA #REQUIRED >

<!--IPOfficeCode
ST.3 code. For example, if the application identification is PCT/IB2013/099999, then
IPOfficeCode value will be "IB" for the International Bureau of WIPO.
-->
<!ELEMENT IPOfficeCode (#PCDATA)>

<!--ApplicationNumberText
The application identification as provided by the office of filing (e.g. PCT/IB2013/099999)
-->
<!ELEMENT ApplicationNumberText (#PCDATA)>

<!--FilingDate
The date of filing of the patent application for which the sequence listing is submitted in
ST.2 format "CCYY-MM-DD", using a 4-digit calendar year, a 2-digit calendar month and a 2-
digit day within the calendar month, e.g., 2015-01-31. For details, please see paragraphs 7
(a) and 11 of WIPO Standard ST.2.
-->
<!ELEMENT FilingDate (#PCDATA)>

******************************************************************************
* INSD Part
******************************************************************************

The purpose of the INSD part of this DTD is to define a customized DTD for sequence
listings to support the work of IP offices while facilitating the data exchange with the
public repositories.

The INSD part is subset of the INSD DTD v1.44 and as such can only be used to generate an
XML instance as it will not support the complete INSD structure.
This part is based on:
The International Nucleotide Sequence Database (INSD) collaboration.

INSDSeq provides the elements of a sequence as presented in the GenBank/EMBL/DDBJ-style flatfile formats. Not all elements are used here.

-->

<!--INSDSeq
Sequence data. Changed INSD V1.5 DTD elements, INSDSeq_division and INSDSeq_sequence from optional to mandatory per business requirements. -->

<!ELEMENT INSDSeq (INSDSeq_length, INSDSeq_moltype, INSDSeq_division, INSDSeq_other-seqids?, INSDSeq_feature-table?, INSDSeq_sequence) >

<!--INSDSeq_length
The length of the sequence. INSDSeq_length allows only integer. -->

<!ELEMENT INSDSeq_length (#PCDATA)>

<!--INSDSeq_moltype
Admissible values: DNA, RNA, AA -->

<!ELEMENT INSDSeq_moltype (#PCDATA)>

<!--INSDSeq_division
Indication that a sequence is related to a patent application. Must be populated with the value PAT. -->

<!ELEMENT INSDSeq_division (#PCDATA)>

<!--INSDSeq_other-seqids
In the context of data exchange with database providers, the Patent Offices should populate for each sequence the element INSDSeq_other-seqids with one INSDSeqid containing a reference to the corresponding published patent and the sequence identification. -->

<!ELEMENT INSDSeq_other-seqids (INSDSeqid?)>

<!--INSDSeq_feature-table
Information on the location and roles of various regions within a particular sequence. Whenever the element INSDSeq_feature-table is used, it must contain at least one feature. -->

<!ELEMENT INSDSeq_feature-table (INSDFeature+)>

<!--INSDSeq_sequence
The residues of the sequence. The sequence must not contain numbers, punctuation or whitespace characters. -->

<!ELEMENT INSDSeq_sequence (#PCDATA)>

<!--INSDSeqid
Intended for the use of Patent Offices in data exchange only.

Format:
pat|{office code}|{publication number}|{document kind code}|{Sequence identification number}

where office code is the code of the IP office publishing the patent document, publication number is the publication number of the application or patent, document kind code is the letter codes to distinguish patent documents as defined in ST.16 and Sequence identification number is the number of the sequence in that application or patent

Example:
pat|WO|2013999999|A1|123456
This represents the 123456th sequence from WO patent publication No. 2013999999 (A1)

<!-- INSDSeqid (#PCDATA) -->

<!-- INSDFeature Description of one feature. -->

<!-- INSDFeature_key A word or abbreviation indicating a feature. -->

<!-- INSDFeature_location Region of the presented sequence which corresponds to the feature. -->

<!-- INSDFeature_quals List of qualifiers containing auxiliary information about a feature. -->

<!-- INSDQualifier Additional information about a feature. For coding sequences and variants see the ST.26 main body document. -->

<!-- INSDQualifier_name Name of the qualifier. -->

<!-- INSDQualifier_value Value of the qualifier. -->

[Annex III follows]
ANNEX III

SEQUENCE LISTING SPECIMEN (XML file)

Version 1.1.1

Adopted Approved by the Committee on WIPO Standards (CWS)
at its reconvened fourth session on March 24, 2016 and at its fifth session on June 2, 2017

<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ST26SequenceListing PUBLIC "-//WIPO//DTD SEQUENCE LISTING 1.0//EN" "ST26SequenceListing_V1_0.dtd">
<ST26SequenceListing
dtdVersion="V1_0"
fileNam="AnnexIII_Sequence_Listing_Specimen.xml"
softwareName="SEQL-software-name" softwareVersion="1.01"
productionDate="2017-06-02">
<ApplicationIdentification>
  <IPOfficeCode>IB</IPOfficeCode>
  <ApplicationNumberText>PCT/IB2015/099999</ApplicationNumberText>
  <FilingDate>2015-01-31</FilingDate>
</ApplicationIdentification>

<EarliestPriorityApplicationIdentification>
  <IPOfficeCode>IB</IPOfficeCode>
  <ApplicationNumberText>PCT/IB2014/111111</ApplicationNumberText>
  <FilingDate>2014-01-30</FilingDate>
</EarliestPriorityApplicationIdentification>

<ApplicantName languageCode="JA" JA>出願製薬株式会社</ApplicantName>
<ApplicantNameLatin>Shutsugan Pharmaceuticals Kabushiki Kaisha</ApplicantNameLatin>

<InventorName languageCode="JA" JA>特許 太郎</InventorName>

<InventionTitle languageCode="JA" JA>タンパク質をコードするマウスabcd-1遺伝子</InventionTitle>

<InventionTitle languageCode="EN" en>Mus musculus abcd-1 gene for efg protein</InventionTitle>

<SequenceTotalQuantity>11</SequenceTotalQuantity>
SEQUENCEDATA sequenceIDNumber="1">
<INSDseq>
  <INSDseq_length>133</INSDseq_length>
  <INSDseq_moltype>DNA</INSDseq_moltype>
  <INSDseq_division>PAT</INSDseq_division>
  <INSDseq_feature-table>
    <INSDFeature>
      <INSDFeature_key>source</INSDFeature_key>
      <INSDFeature_location>1..133</INSDFeature_location>
      <INSDFeature_quals>
        <INSDQualifier>
          <INSDQualifier_name>organism</INSDQualifier_name>
          <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
        </INSDQualifier>
        <INSDQualifier>
          <INSDQualifier_name>mol_type</INSDQualifier_name>
          <INSDQualifier_value>genomic DNA</INSDQualifier_value>
        </INSDQualifier>
      </INSDFeature_quals>
    </INSDFeature>
  </INSDseq_feature-table>
</INSDseq>
<INSDSeq_sequence>
ataagattaaacataaaaggtgataaattgagatttgataaaaaagtttaggttagttgagagagatatttgaga
cggcatggagagacagagttaaaggacattaaaagggcatatatgacatata</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="2">
<INSDSeq>
<INSDSeq_length>29</INSDSeq_length>
<INSDSeq_moltype>AA</INSDSeq_moltype>
<INSDSeq_division>PA</INSDSeq_division>
<INSDSeq_feature-table>
<INSDFeature>
<INSDFeature_key>SOURCE</INSDFeature_key>
<INSDFeature_location>1..29</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>ORGANISM</INSDQualifier_name>
<INSDQualifier_value>synthetic construct</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
<INSDQualifier_value>protein</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>NOTE</INSDQualifier_name>
<INSDQualifier_value>Synthetic peptide antigen fragment</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
</INSDSeq_feature-table>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="3">
<INSDSeq>
<INSDSeq_length>62</INSDSeq_length>
<INSDSeq_moltype>DNA</INSDSeq_moltype>
<INSDSeq_division>PA</INSDSeq_division>
<INSDSeq_feature-table>
<INSDFeature>
<INSDFeature_key>source</INSDFeature_key>
<INSDFeature_location>1..62</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>organism</INSDQualifier_name>
<INSDQualifier_value>Homo sapiens</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>mol_type</INSDQualifier_name>
<INSDQualifier_value>genomic DNA</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
</INSDSeq_feature-table>
</INSDSeq>
</SequenceData>

<INSDSeq_sequence>
GSLSDVRKDVEKRIDKALEAFKNKMDKE</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<INSDQualifier_name>translation</INSDQualifier_name>
<INSDQualifier_value>MLAPDCPFDPTRIYSSSLC</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>protein_id</INSDQualifier_name>
<INSDQualifier_value>4</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
</INSDSeq_feature-table>
<INSDFeature>
<INSDFeature_key>SOURCE</INSDFeature_key>
<INSDFeature_location>1..19</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>ORGANISM</INSDQualifier_name>
<INSDQualifier_value>Homo sapiens</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
<INSDQualifier_value>protein</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
</INSDSeq_feature-table>
<INSDFeature>
<INSDFeature_key>source</INSDFeature_key>
<INSDFeature_location>1..133</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>organism</INSDQualifier_name>
<INSDQualifier_value>Solanum lycopersicum</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>mol_type</INSDQualifier_name>
<INSDQualifier_value>genomic DNA</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>note</INSDQualifier_name>
<INSDQualifier_value>common name: tomato</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
</INSDSeq>
</SequenceData>
</SequenceData>
<SequenceData_sequenceIDNumber="5">
<INSDSeq>
<INSDSeq_length>133</INSDSeq_length>
<INSDSeq_moltype>DNA</INSDSeq_moltype>
<INSDSeq_division>PAT</INSDSeq_division>
<INSDSeq_feature-table>
<INSDFeature>
<INSDFeature_key>source</INSDFeature_key>
<INSDFeature_location>1..133</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>organism</INSDQualifier_name>
<INSDQualifier_value>Solanum lycopersicum</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>mol_type</INSDQualifier_name>
<INSDQualifier_value>genomic DNA</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>note</INSDQualifier_name>
<INSDQualifier_value>common name: tomato</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
</INSDSeq_feature-table>
</INSDSeq>
</SequenceData>
</SequenceData>
<SequenceData_sequenceIDNumber="4">
<INSDSeq>
<INSDSeq_length>19</INSDSeq_length>
<INSDSeq_moltype>AA</INSDSeq_moltype>
<INSDSeq_division>PAT</INSDSeq_division>
<INSDSeq_feature-table>
<INSDFeature>
<INSDFeature_key>SOURCE</INSDFeature_key>
<INSDFeature_location>1..19</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>ORGANISM</INSDQualifier_name>
<INSDQualifier_value>Homo sapiens</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
<INSDQualifier_value>protein</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
</INSDSeq_feature-table>
</INSDSeq>
</SequenceData>
</SequenceData>
<SequenceData_sequenceIDNumber="5">
<INSDSeq>
<INSDSeq_length>133</INSDSeq_length>
<INSDSeq_moltype>DNA</INSDSeq_moltype>
<INSDSeq_division>PAT</INSDSeq_division>
<INSDSeq_feature-table>
<INSDFeature>
<INSDFeature_key>source</INSDFeature_key>
<INSDFeature_location>1..133</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>organism</INSDQualifier_name>
<INSDQualifier_value>Solanum lycopersicum</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>mol_type</INSDQualifier_name>
<INSDQualifier_value>genomic DNA</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>note</INSDQualifier_name>
<INSDQualifier_value>common name: tomato</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
</INSDSeq_feature-table>
</INSDSeq>
</SequenceData>
</SequenceData>
<SequenceData_sequenceIDNumber="4">
<INSDSeq>
<INSDSeq_length>19</INSDSeq_length>
<INSDSeq_moltype>AA</INSDSeq_moltype>
<INSDSeq_division>PAT</INSDSeq_division>
<INSDSeq_feature-table>
<INSDFeature>
<INSDFeature_key>SOURCE</INSDFeature_key>
<INSDFeature_location>1..19</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>ORGANISM</INSDQualifier_name>
<INSDQualifier_value>Homo sapiens</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
<INSDQualifier_value>protein</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
</INSDSeq_feature-table>
</INSDSeq>
</SequenceData>
</SequenceData>
<SequenceData_sequenceIDNumber="5">
<INSDSeq>
<INSDSeq_length>133</INSDSeq_length>
<INSDSeq_moltype>DNA</INSDSeq_moltype>
<INSDSeq_division>PAT</INSDSeq_division>
<INSDSeq_feature-table>
<INSDFeature>
<INSDFeature_key>source</INSDFeature_key>
<INSDFeature_location>1..133</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>organism</INSDQualifier_name>
<INSDQualifier_value>Solanum lycopersicum</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>mol_type</INSDQualifier_name>
<INSDQualifier_value>genomic DNA</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>note</INSDQualifier_name>
<INSDQualifier_value>common name: tomato</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
</INSDSeq_feature-table>
</INSDSeq>
</SequenceData>
</SequenceData>
<INSDSeq_feature-table>
  <INSDFeature>
    <INSDFeature_key>variation</INSDFeature_key>
    <INSDFeature_location>60</INSDFeature_location>
    <INSDFeature_quals>
      <INSDQualifier>
        <INSDQualifier_name>replace</INSDQualifier_name>
        <INSDQualifier_value>c</INSDQualifier_value>
      </INSDQualifier>
    </INSDFeature_quals>
  </INSDFeature>
</INSDSeq_feature-table>

<INSDSeq_sequence>
atgaatataacacaaaaagagtattgggattaaatatgttagttgctgaggaaggatattttagggcattgagagagacaggcattataagattgctgatgacaaat</INSDSeq_sequence>
</INSDSeq>

<SequenceData sequenceIDNumber="6">
  <INSDSeq>
    <INSDSeq_length>29</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..29</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>synthetic construct</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
    </INSDSeq_feature-table>
  </INSDSeq>
</SequenceData>
<INSDQualifier_value>protein</INSDQualifier_value></INSDQualifier>
</INSDFeature>
<INSDFeature>
<INSDFeature_key>MOD_RES</INSDFeature_key>
<INSDFeature_location>3</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>NOTE</INSDQualifier_name>
<INSDQualifier_value>N-acetylated</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
<INSDFeature>
<INSDFeature_key>SITE</INSDFeature_key>
<INSDFeature_location>7</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>NOTE</INSDQualifier_name>
<INSDQualifier_value>Orn</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
<INSDFeature>
<INSDFeature_key>SITE</INSDFeature_key>
<INSDFeature_location>13</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>NOTE</INSDQualifier_name>
<INSDQualifier_value>D-Arginine</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
<INSDFeature>
<INSDFeature_key>UNSURE</INSDFeature_key>
<INSDFeature_location>15</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>NOTE</INSDQualifier_name>
<INSDQualifier_value>A or V</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
<INSDFeature>
<INSDFeature_key>VARIANT</INSDFeature_key>
<INSDFeature_location>20</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>NOTE</INSDQualifier_name>
<INSDQualifier_value>I, A, F, Y, sIle, MeIle, or Nle</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
<INSDQualifier>
  <INSQualified_name>organism</INSQualified_name>
  <INSQualified_value>Candida albicans</INSQualified_value>
</INSDQualifier>

<INSDQualifier>
  <INSQualified_name>mol_type</INSQualified_name>
  <INSQualified_value>genomic DNA</INSQualified_value>
</INSDQualifier>

<INSDSeq>
  <INSSeq_length>288</INSSeq_length>
  <INSSeq_moltype>DNA</INSSeq_moltype>
  <INSSeq_division>PAT</INSSeq_division>
  <INSSeq_feature-table>
    <INSDFeature>
      <INSDFeature_key>source</INSDFeature_key>
      <INSDFeature_location>1..288</INSDFeature_location>
      <INSDFeature_quals>
        <INSDQualifier>
          <INSQualified_name>organism</INSQualified_name>
          <INSQualified_value>Candida albicans</INSQualified_value>
        </INSDQualifier>
        <INSDQualifier>
          <INSQualified_name>mol_type</INSQualified_name>
          <INSQualified_value>genomic DNA</INSQualified_value>
        </INSDQualifier>
      </INSDFeature_quals>
    </INSDFeature>
  </INSSeq_feature-table>
</INSDSeq>

<INSDSeq_sequence>
cgacccacgcgtccgaggaaccaaccatcacgtttgaggacttcgtgaaggaattggataatacccgtccctaccaaaatg
cgagcgccgactcattgctcctcgtaccgtcgagcgg
</INSDSeq_sequence>
<INSDFeature_table>
<INSDFeature>
<INSDFeature_key>SOURCE</INSDFeature_key>
<INSDFeature_location>1..95</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>ORGANISM</INSDQualifier_name>
<INSDQualifier_value>Candida albicans</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
</INSDFeature_table>
</INSDSeq>
</SequenceData>
</ST26SequenceListing>

[Annex IV follows]
ANNEX IV

CHARACTER SUBSET FROM THE UNICODE BASIC LATIN CODE TABLE

Version 1.1

Adopted by the Committee on WIPO Standards (CWS) at its reconvened fourth session on March 24, 2016 and fifth session on June 2, 2017.

The ampersand character (0026) is only permitted as part of a predefined entity or as part of a numeric character reference (&#nnnn). The quotation mark (0022), the apostrophe (0027), the less-than sign (003C), and the greater-than sign (003E) are not permitted and must be represented by their predefined entities.

<table>
<thead>
<tr>
<th>Unicode code point</th>
<th>Character</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0020</td>
<td></td>
<td>SPACE</td>
</tr>
<tr>
<td>0021</td>
<td>!</td>
<td>EXCLAMATION MARK</td>
</tr>
<tr>
<td>0023</td>
<td>#</td>
<td>NUMBER SIGN</td>
</tr>
<tr>
<td>0024</td>
<td>$</td>
<td>DOLLAR SIGN</td>
</tr>
<tr>
<td>0025</td>
<td>%</td>
<td>PERCENT SIGN</td>
</tr>
<tr>
<td>0026</td>
<td>&amp;</td>
<td>AMPERSAND</td>
</tr>
<tr>
<td>0027</td>
<td>(</td>
<td>LEFT PARENTHESIS</td>
</tr>
<tr>
<td>0028</td>
<td>)</td>
<td>RIGHT PARENTHESIS</td>
</tr>
<tr>
<td>0029</td>
<td>,</td>
<td>COMMA</td>
</tr>
<tr>
<td>002A</td>
<td>+</td>
<td>PLUS SIGN</td>
</tr>
<tr>
<td>002B</td>
<td>-</td>
<td>HYPHEN-MINUS</td>
</tr>
<tr>
<td>002C</td>
<td>.</td>
<td>FULL STOP</td>
</tr>
<tr>
<td>002D</td>
<td>/</td>
<td>SOLIDUS</td>
</tr>
<tr>
<td>0030</td>
<td>0</td>
<td>DIGIT ZERO</td>
</tr>
<tr>
<td>0031</td>
<td>1</td>
<td>DIGIT ONE</td>
</tr>
<tr>
<td>0032</td>
<td>2</td>
<td>DIGIT TWO</td>
</tr>
<tr>
<td>0033</td>
<td>3</td>
<td>DIGIT THREE</td>
</tr>
<tr>
<td>0034</td>
<td>4</td>
<td>DIGIT FOUR</td>
</tr>
<tr>
<td>0035</td>
<td>5</td>
<td>DIGIT FIVE</td>
</tr>
<tr>
<td>0036</td>
<td>6</td>
<td>DIGIT SIX</td>
</tr>
<tr>
<td>0037</td>
<td>7</td>
<td>DIGIT SEVEN</td>
</tr>
<tr>
<td>0038</td>
<td>8</td>
<td>DIGIT EIGHT</td>
</tr>
<tr>
<td>0039</td>
<td>9</td>
<td>DIGIT NINE</td>
</tr>
<tr>
<td>003A</td>
<td>:</td>
<td>COLON</td>
</tr>
<tr>
<td>003B</td>
<td>;</td>
<td>SEMICOLON</td>
</tr>
<tr>
<td>003C</td>
<td>=</td>
<td>EQUALS SIGN</td>
</tr>
<tr>
<td>003D</td>
<td>?</td>
<td>QUESTION MARK</td>
</tr>
<tr>
<td>003E</td>
<td>@</td>
<td>COMMERCIAL AT</td>
</tr>
<tr>
<td>0041</td>
<td>A</td>
<td>LATIN CAPITAL LETTER A</td>
</tr>
<tr>
<td>0042</td>
<td>B</td>
<td>LATIN CAPITAL LETTER B</td>
</tr>
<tr>
<td>0043</td>
<td>C</td>
<td>LATIN CAPITAL LETTER C</td>
</tr>
<tr>
<td>0044</td>
<td>D</td>
<td>LATIN CAPITAL LETTER D</td>
</tr>
<tr>
<td>0045</td>
<td>E</td>
<td>LATIN CAPITAL LETTER E</td>
</tr>
<tr>
<td>0046</td>
<td>F</td>
<td>LATIN CAPITAL LETTER F</td>
</tr>
<tr>
<td>0047</td>
<td>G</td>
<td>LATIN CAPITAL LETTER G</td>
</tr>
<tr>
<td>0048</td>
<td>H</td>
<td>LATIN CAPITAL LETTER H</td>
</tr>
<tr>
<td>0049</td>
<td>I</td>
<td>LATIN CAPITAL LETTER I</td>
</tr>
<tr>
<td>004A</td>
<td>J</td>
<td>LATIN CAPITAL LETTER J</td>
</tr>
<tr>
<td>004B</td>
<td>K</td>
<td>LATIN CAPITAL LETTER K</td>
</tr>
<tr>
<td>004C</td>
<td>L</td>
<td>LATIN CAPITAL LETTER L</td>
</tr>
<tr>
<td>004D</td>
<td>M</td>
<td>LATIN CAPITAL LETTER M</td>
</tr>
<tr>
<td>004E</td>
<td>N</td>
<td>LATIN CAPITAL LETTER N</td>
</tr>
<tr>
<td>004F</td>
<td>O</td>
<td>LATIN CAPITAL LETTER O</td>
</tr>
<tr>
<td>Unicode code point</td>
<td>Character</td>
<td>Name</td>
</tr>
<tr>
<td>--------------------</td>
<td>----------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td>0050</td>
<td>P</td>
<td>LATIN CAPITAL LETTER P</td>
</tr>
<tr>
<td>0051</td>
<td>Q</td>
<td>LATIN CAPITAL LETTER Q</td>
</tr>
<tr>
<td>0052</td>
<td>R</td>
<td>LATIN CAPITAL LETTER R</td>
</tr>
<tr>
<td>0053</td>
<td>S</td>
<td>LATIN CAPITAL LETTER S</td>
</tr>
<tr>
<td>0054</td>
<td>T</td>
<td>LATIN CAPITAL LETTER T</td>
</tr>
<tr>
<td>0055</td>
<td>U</td>
<td>LATIN CAPITAL LETTER U</td>
</tr>
<tr>
<td>0056</td>
<td>V</td>
<td>LATIN CAPITAL LETTER V</td>
</tr>
<tr>
<td>0057</td>
<td>W</td>
<td>LATIN CAPITAL LETTER W</td>
</tr>
<tr>
<td>0058</td>
<td>X</td>
<td>LATIN CAPITAL LETTER X</td>
</tr>
<tr>
<td>0059</td>
<td>Y</td>
<td>LATIN CAPITAL LETTER Y</td>
</tr>
<tr>
<td>005A</td>
<td>Z</td>
<td>LATIN CAPITAL LETTER Z</td>
</tr>
<tr>
<td>005B</td>
<td>[</td>
<td>LEFT SQUARE BRACKET</td>
</tr>
<tr>
<td>005C</td>
<td>\</td>
<td>REVERSE SOLIDUS</td>
</tr>
<tr>
<td>005D</td>
<td>]</td>
<td>RIGHT SQUARE BRACKET</td>
</tr>
<tr>
<td>005E</td>
<td>^</td>
<td>CIRCUMFLEX ACCENT</td>
</tr>
<tr>
<td>005F</td>
<td>_</td>
<td>LOW LINE</td>
</tr>
<tr>
<td>0060</td>
<td>`</td>
<td>GRAVE ACCENT</td>
</tr>
<tr>
<td>0061</td>
<td>a</td>
<td>LATIN SMALL LETTER A</td>
</tr>
<tr>
<td>0062</td>
<td>b</td>
<td>LATIN SMALL LETTER B</td>
</tr>
<tr>
<td>0063</td>
<td>c</td>
<td>LATIN SMALL LETTER C</td>
</tr>
<tr>
<td>0064</td>
<td>d</td>
<td>LATIN SMALL LETTER D</td>
</tr>
<tr>
<td>0065</td>
<td>e</td>
<td>LATIN SMALL LETTER E</td>
</tr>
<tr>
<td>0066</td>
<td>f</td>
<td>LATIN SMALL LETTER F</td>
</tr>
<tr>
<td>0067</td>
<td>g</td>
<td>LATIN SMALL LETTER G</td>
</tr>
<tr>
<td>0068</td>
<td>h</td>
<td>LATIN SMALL LETTER H</td>
</tr>
<tr>
<td>0069</td>
<td>i</td>
<td>LATIN SMALL LETTER I</td>
</tr>
<tr>
<td>006A</td>
<td>j</td>
<td>LATIN SMALL LETTER J</td>
</tr>
<tr>
<td>006B</td>
<td>k</td>
<td>LATIN SMALL LETTER K</td>
</tr>
<tr>
<td>006C</td>
<td>l</td>
<td>LATIN SMALL LETTER L</td>
</tr>
<tr>
<td>006D</td>
<td>m</td>
<td>LATIN SMALL LETTER M</td>
</tr>
<tr>
<td>006E</td>
<td>n</td>
<td>LATIN SMALL LETTER N</td>
</tr>
<tr>
<td>006F</td>
<td>o</td>
<td>LATIN SMALL LETTER O</td>
</tr>
<tr>
<td>0070</td>
<td>p</td>
<td>LATIN SMALL LETTER P</td>
</tr>
<tr>
<td>0071</td>
<td>q</td>
<td>LATIN SMALL LETTER Q</td>
</tr>
<tr>
<td>0072</td>
<td>r</td>
<td>LATIN SMALL LETTER R</td>
</tr>
<tr>
<td>0073</td>
<td>s</td>
<td>LATIN SMALL LETTER S</td>
</tr>
<tr>
<td>0074</td>
<td>t</td>
<td>LATIN SMALL LETTER T</td>
</tr>
<tr>
<td>0075</td>
<td>u</td>
<td>LATIN SMALL LETTER U</td>
</tr>
<tr>
<td>0076</td>
<td>v</td>
<td>LATIN SMALL LETTER V</td>
</tr>
<tr>
<td>0077</td>
<td>w</td>
<td>LATIN SMALL LETTER W</td>
</tr>
<tr>
<td>0078</td>
<td>x</td>
<td>LATIN SMALL LETTER X</td>
</tr>
<tr>
<td>0079</td>
<td>y</td>
<td>LATIN SMALL LETTER Y</td>
</tr>
<tr>
<td>007A</td>
<td>z</td>
<td>LATIN SMALL LETTER Z</td>
</tr>
<tr>
<td>007B</td>
<td>{</td>
<td>LEFT CURLY BRACKET</td>
</tr>
<tr>
<td>007C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>007D</td>
<td>}</td>
<td>RIGHT CURLY BRACKET</td>
</tr>
<tr>
<td>007E</td>
<td>~</td>
<td>TILDE</td>
</tr>
</tbody>
</table>

[Annex V follows]
ANNEX V

ADDITIONAL DATA EXCHANGE REQUIREMENTS (FOR PATENT OFFICES ONLY)

Version 1.1.1

Adopted Approved by the Committee on WIPO Standards (CWS)
at its reconvened fourth session on March 24, 2016 fifth session on June 2, 2017

In the context of data exchange with database providers (INSD members), the Patent Offices should populate for each sequence the element INSDSeq_other-seqids with one INSDSeqid containing a reference to the corresponding published patent and the sequence identification number in the following format:

\[ \text{pat|office code|publication number|document kind code|sequence identification number} \]

where office code is the code of the IP office publishing the patent document as set forth in ST.3; document kind code is the code for the identification of different kinds of patent documents as set forth in ST.16; publication number is the publication number of the application or patent; and Sequence identification number is the number of the sequence in that application or patent.

Example:

\[ \text{pat|WO|2013999999|A1|123456} \]

Which would be translated into a valid XML instance as:

\[
\begin{align*}
&<\text{INSDSeq_other-seqids}>
&< \text{INSDSeqid}>\text{pat|WO|2013999999|A1|123456}</\text{INSDSeqid}>
&</\text{INSDSeq_other-seqids}>
\end{align*}
\]

Where “123456” is the 123456th sequence from the WO publication no. 2013999999 (A1).

[Annex VI follows]
INTRODUCTION

This Standard indicates as one of its purposes, to “allow applicants to draw up a single sequence listing in a patent application acceptable for the purposes of both international and national or regional procedures.” The purpose of this Guidance Document is to ensure that all applicants and Intellectual Property Offices (IPOs) understand and agree on the requirements for inclusion and representation of sequence disclosures, such that this purpose is realized.

This guidance document consists of this introduction, an example index, examples of sequence disclosures, and an appendix containing a sequence listing in XML with sequences from the examples. This introduction explains certain concepts and terminology used in the remainder of this document. The examples illustrate the requirements of specific paragraphs of the standard and each example has been designated with the most relevant paragraph number. Some examples further illustrate other paragraphs and appropriate cross-references are indicated at the end of each example.

The index provides page numbers for the examples and any indicated cross-references. Each sequence in an example that either must or may be included in a sequence listing has been assigned a sequence identification number (SEQ ID NO) and appears in XML format in the Appendix to this document.

For each example, any explanatory information presented with a sequence is intended to be considered as the entirety of the disclosure concerning that sequence. The given answers take into account only the information explicitly presented in the example.

The guidance provided in this document is directed to the preparation of a sequence listing for provision on the filing date of a patent application. Preparation of a sequence listing for provision subsequent to the filing date of a patent application must take into account whether the information provided could be considered by an IPO to add subject matter to the original disclosure. Therefore, it is possible that the guidance provided in this document may not be applicable to a sequence listing provided subsequent to the filing date of a patent application.

Preparation of a sequence listing

Sequence listing preparation for a patent application requires consideration of the following questions:

1. Does ST.26 paragraph 7 require inclusion of a particular disclosed sequence?
2. If inclusion of a particular disclosed sequence is not required, is inclusion of that sequence permitted by ST.26?
3. If inclusion of a particular disclosed sequence is required or permitted by ST.26, how should that sequence be represented in the sequence listing?

Regarding the first question, ST.26 paragraph 7 (with certain restrictions) requires inclusion of a sequence disclosed in a patent application by enumeration of its residues, where the sequence contains ten or more specifically defined nucleotides or four or more specifically defined amino acids.

Regarding the second question, ST.26 paragraph 8 prohibits inclusion of any sequences having fewer than ten specifically defined nucleotides or four specifically defined amino acids.

A clear understanding of “enumeration of its residues” and “specifically defined” is necessary to answer these two questions.

Regarding the third question, this document provides sequence disclosures which exemplify a variety of scenarios together with a complete discussion of the preferred means of representation of each sequence, or where a sequence contains multiple variations - the “most encompassing sequence”, in accordance with this Standard. Since it is impossible to address every possible unusual sequence scenario, this guidance document attempts to set forth the reasoning behind the approach to each example and the manner in which ST.26 provisions are applied, such that the same reasoning can be applied to other sequence scenarios not exemplified.
Enumeration of its residues

ST.26 paragraph 3(c) defines “enumeration of its residues” as disclosure of a sequence in a patent application by listing, in order, each residue of the sequence, wherein (i) the residue is represented by a name, abbreviation, symbol, or structure; or (ii) multiple residues are represented by a shorthand formula. A sequence should be disclosed in a patent application by ‘enumeration of its residues’ using conventional symbols, which are the nucleotide symbols set forth in Section 1, Table 1 of ST.26 Annex 1 (i.e. the lower case symbols or their upper case equivalents) and the amino acid symbols set forth in Section 3, Table 3 of ST.26 Annex 1 (i.e. the upper case symbols or their lower case equivalents). Symbols other than those set forth in these tables are “nonconventional”.

A sequence is sometimes disclosed in a non-preferred manner by "enumeration of its residues" using conventional abbreviations or full names (as opposed to conventional symbols) as set forth in Tables A and B below, conventional symbols or abbreviations used in a nonconventional manner, nonconventional symbols or abbreviations, chemical formulas/structures, or shorthand formulas. Care should be taken to disclose sequences in the preferred manner; however, where sequences are disclosed in a non-preferred manner, consultation of the explanation of the sequence in the disclosure may be necessary to determine the meaning of the non-preferred symbol or abbreviation.

Where a conventional symbol or abbreviation is used, the explanation of the sequence in the disclosure must still be consulted to confirm that the symbol is used in a conventional manner. Otherwise, if the symbol is used in a nonconventional manner, the explanation is necessary to determine whether ST.26 paragraph 7 requires inclusion in the sequence listing or whether paragraph 8 prohibits inclusion.

Where a nonconventional symbol or abbreviation is disclosed as equivalent to a conventional symbol or abbreviation (e.g., “Z” means “A”), or to a specific sequence of conventional symbols (e.g., “Zn” means “agga”), then the sequence is interpreted as though it were disclosed using the equivalent conventional symbol(s) or abbreviation(s), to determine whether ST.26 paragraph 7 requires inclusion in the sequence listing or whether paragraph 8 prohibits inclusion. Where a nonconventional nucleotide symbol is used as an ambiguity symbol (e.g., X1 = inosine or pseudouridine), but is not equivalent to one of the conventional ambiguity symbols in Section 1, Table 1 (i.e., “m”, “r”, “w”, “s”, “y”, “k”, “v”, “h”, “d”, “b”, or “n”), then the residue is interpreted as an “n” residue to determine whether ST.26 Paragraph 7 requires inclusion of the sequence in the sequence listing or whether ST.26 Paragraph 8 prohibits inclusion. Similarly, where a nonconventional amino acid symbol is used as an ambiguity symbol (e.g., “Z” means “A”, “G”, “S” or “T”), but is not equivalent to one of the conventional ambiguity symbols in Section 3, Table 3 (i.e., B, Z, J, or X), then the residue is interpreted as an “X” residue to determine whether ST.26 paragraph 7 requires inclusion of the sequence in the sequence listing or whether ST.26 paragraph 8 prohibits inclusion.

Specifically defined

ST.26 paragraph 3(k) defines “specifically defined” as any nucleotide other than those represented by the symbol “n” and any amino acid other than those represented by the symbol “X”, listed in Annex I, wherein “n” and “X” are used in a conventional manner as described in Section 1, Table 1 (i.e., “a or c or g or t or u; ‘unknown’ or ‘other’”) and Section 3, Table 3 (i.e., A or R or N or D or C or Q or E or G or H or I or L or K or M or F or P or O or S or U or T or W or Y or V; ‘unknown’ or ‘other’), respectively. The discussion above concerning conventional symbols or nonconventional symbols or abbreviations and their use in a conventional or nonconventional manner will be taken into account to determine whether a nucleotide or an amino acid is “specifically defined”.

Most encompassing sequence

Where a sequence that meets the requirements of paragraph 7 is disclosed by enumeration of its residues only once in an application, but is described differently in multiple embodiments, e.g. in one embodiment “X” in one or more locations could be any amino acid, but in further embodiments, “X” could be only a limited number of amino acids, ST.26 requires inclusion in a sequence listing of only the single sequence that has been enumerated by its residues. As per paragraphs 15 and 27, where such a sequence contains multiple “n” or “X” ambiguity symbols, “n” or “X” is construed to represent any nucleotide or amino acid, respectively, in the absence of further annotation. Consequently, the single sequence required to be included is the most encompassing sequence disclosed. The most encompassing sequence is the single sequence having variant residues which are represented by the most restrictive ambiguity symbols that include the most disclosed embodiments. However, inclusion of additional specific sequences is strongly encouraged where practical, e.g. which represent additional embodiments that are a key part of the invention. Inclusion of the additional sequences allows for a more thorough search and provides public notice of the subject matter for which a patent is sought.

1 NOTE: While an application disclosure may represent nucleotides or amino acids with either lower case or upper case symbols, for a sequence included in a sequence listing, only lower case letters must be used for representation of a nucleotide sequence (see ST.26 paragraph 13) and only upper case letters must be used for representation of an amino acid sequence (see ST.26 paragraph 26).
Usage of Ambiguity Symbol

Proper Usage of the Ambiguity Symbol “n” in a Sequence Listing

The symbol “n”:

a. must not be used to represent anything other than a single nucleotide;

b. will be construed as any one of “a”, “c”, “g”, or “t/u” except where it is used with a further description;

c. should be used to represent any of the following nucleotides together with a further description:
   i. modified nucleotide, e.g., natural, synthetic, or non-naturally occurring, that cannot otherwise be represented by any other symbol in Annex I (see Section 1, Table 1);
   ii. “unknown” nucleotide, i.e., not determined, not disclosed, or unsure;
   iii. an abasic site; or

d. may be used to represent a sequence variant, i.e., alternatives, deletions, insertions, or substitutions, where “n” is the most restrictive ambiguity symbol.

Proper Usage of the Ambiguity Symbol “X” in a Sequence Listing

The symbol “X”:

a. must not be used to represent anything other than a single amino acid;


c. should be used to represent any of the following amino acids together with a further description:
   i. modified amino acid, e.g., natural, synthetic, or non-naturally occurring, that cannot otherwise be represented by any other symbol in Annex I (see Section 3, Table 3);
   ii. “unknown” amino acid, i.e., not determined, not disclosed, or unsure;


d. may be used to represent a sequence variant, i.e., alternatives, deletions, insertions, or substitutions, where “X” is the most restrictive ambiguity symbol.
Table A – Conventional Nucleotide Symbols, Abbreviations, and Names

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Abbreviation</th>
<th>Nucleotide Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>Adenine</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>Cytosine</td>
</tr>
<tr>
<td>G</td>
<td></td>
<td>Guanine</td>
</tr>
<tr>
<td>T</td>
<td>Thymine in DNA, Uracil in RNA (t/u)</td>
<td></td>
</tr>
<tr>
<td>m</td>
<td>a or c</td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>a or g</td>
<td></td>
</tr>
<tr>
<td>w</td>
<td>a or t/u</td>
<td></td>
</tr>
<tr>
<td>s</td>
<td>c or g</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>c or t/u</td>
<td></td>
</tr>
<tr>
<td>k</td>
<td>g or t/u</td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>a or c or g; not t/u</td>
<td></td>
</tr>
<tr>
<td>h</td>
<td>a or c or t/u; not g</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>a or g or t/u; not c</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>c or g or t/u; not a</td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>a or c or g or t/u; “unknown” or “other”</td>
<td></td>
</tr>
</tbody>
</table>
### Table B – Conventional Amino Acid Symbols, Abbreviations, and Names

<table>
<thead>
<tr>
<th>Symbol</th>
<th>3-Letter Abbreviation</th>
<th>Amino Acid Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Ala</td>
<td>Alanine</td>
</tr>
<tr>
<td>R</td>
<td>Arg</td>
<td>Arginine</td>
</tr>
<tr>
<td>N</td>
<td>Asn</td>
<td>Asparagine</td>
</tr>
<tr>
<td>D</td>
<td>Asp</td>
<td>Aspartic Acid (Aspartate)</td>
</tr>
<tr>
<td>C</td>
<td>Cys</td>
<td>Cysteine</td>
</tr>
<tr>
<td>E</td>
<td>Glu</td>
<td>Glutamic Acid (Glutamate)</td>
</tr>
<tr>
<td>Q</td>
<td>Gln</td>
<td>Glutamine</td>
</tr>
<tr>
<td>G</td>
<td>Gly</td>
<td>Glycine</td>
</tr>
<tr>
<td>H</td>
<td>His</td>
<td>Histidine</td>
</tr>
<tr>
<td>I</td>
<td>Ile</td>
<td>Isoleucine</td>
</tr>
<tr>
<td>L</td>
<td>Leu</td>
<td>Leucine</td>
</tr>
<tr>
<td>K</td>
<td>Lys</td>
<td>Lysine</td>
</tr>
<tr>
<td>M</td>
<td>Met</td>
<td>Methionine</td>
</tr>
<tr>
<td>E</td>
<td>Phe</td>
<td>Phenylalanine</td>
</tr>
<tr>
<td>P</td>
<td>Pro</td>
<td>Proline</td>
</tr>
<tr>
<td>O</td>
<td>Pyl</td>
<td>Pyrolysine</td>
</tr>
<tr>
<td>S</td>
<td>Ser</td>
<td>Serine</td>
</tr>
<tr>
<td>U</td>
<td>Sec</td>
<td>Selenocysteine</td>
</tr>
<tr>
<td>T</td>
<td>Thr</td>
<td>Threonine</td>
</tr>
<tr>
<td>W</td>
<td>Trp</td>
<td>Tryptophan</td>
</tr>
<tr>
<td>Y</td>
<td>Tyr</td>
<td>Tyrosine</td>
</tr>
<tr>
<td>V</td>
<td>Val</td>
<td>Valine</td>
</tr>
<tr>
<td>B</td>
<td>Asx</td>
<td>Aspartic acid or Asparagine</td>
</tr>
<tr>
<td>Z</td>
<td>Gix</td>
<td>Glutamine or Glutamic Acid</td>
</tr>
<tr>
<td>J</td>
<td>Xle</td>
<td>Leucine or Isoleucine</td>
</tr>
<tr>
<td>X</td>
<td>Xaa</td>
<td>A or R or N or D or C or Q or E or G or H or I or L or K or M or F or P or O or S or U or T or W or Y or V or “unknown” or “other”</td>
</tr>
</tbody>
</table>
EXAMPLE INDEX

**Paragraph 3(a) – Definition of “amino acid”**

**Example 3(a)-1: D amino acids** ................................................................. 14

*Cros-referenced examples*

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid ........................................ 47
Example 30-1: Feature key “CARBOHYD” .................................................................................. 48

**Paragraph 3(c) – Definition of “enumeration of its residues”**

**Example 3(c)-1: Enumeration of amino acids by chemical structure** ......................................................... 15
**Example 3(c)-2: Shorthand formula for an amino acid sequence** ............................................................ 16

*Cros-referenced examples*

Example 27-1: Shorthand formula for a nucleotide sequence ................................................................. 42
Example 27-3: Shorthand formula - four or more specifically defined amino acids ................................. 44

**Paragraph 3(f) – Definition of “modified nucleotide”**

*Cros-referenced examples*

Example 3(g)-4: Nucleic Acid Analogues ......................................................................................... 20

**Paragraph 3(g) – Definition of “nucleotide”**

**Example 3(g)-1: Nucleotide sequence interrupted by a C3 spacer** ......................................................... 17
**Example 3(g)-2: Nucleotide sequence with residue alternatives, including a C3 spacer** ......................... 18
**Example 3(g)-3: Abasic site** .............................................................................................................. 19
**Example 3(g)-4: Nucleic Acid Analogues** .......................................................................................... 20

*Cros-referenced examples*

Example 11(b)-1: Double-stranded nucleotide sequence – different lengths ........................................ 38
Example 14-1: The symbol “t” represents uracil in RNA ................................................................. 40

**Paragraph 3(k) – Definition of “specifically defined”**

**Example 3(k)-1: Nucleotide ambiguity symbols** .................................................................................. 21
**Example 3(k)-2: Ambiguity symbol “n” used in both a conventional and nonconventional manner** .......... 22
**Example 3(k)-3: Ambiguity symbol “n” used in a nonconventional manner** .......................................... 23
**Example 3(k)-4: Ambiguity symbols other than “n” are “specifically defined”** ........................................ 24
**Example 3(k)-5: Ambiguity abbreviation “Xaa” used in a nonconventional manner** ............................ 25

**Paragraph 7 – Sequences for which inclusion in a sequence listing is required**

*Cros-referenced examples*

**Example 28-1: Encoding nucleotide sequence and encoded amino acid sequence** ................................ 45
Paragraph 7(a) – Nucleotide sequences required in a sequence listing

Example 7(a)-1: Branched nucleotide sequence ................................................................. 26
Example 7(a)-2: Linear nucleotide sequence having a secondary structure .................. 28
Example 7(a)-3: Nucleotide ambiguity symbols used in a nonconventional manner .......... 29
Example 7(a)-4: Nucleotide ambiguity symbols used in a nonconventional manner .......... 30
Example 7(a)-5: Nonconventional nucleotide symbols .................................................. 31
Example 7(a)-6: Nonconventional nucleotide symbols .................................................. 32

Cross-referenced examples
Example 3(p)-1: Nucleotide sequence interrupted by a C3 spacer ................................. 17
Example 3(p)-2: Nucleotide sequence with residue alternatives, including a C3 spacer ....... 18
Example 3(p)-3: Abasic site ................................................................................................. 19
Example 3(p)-4: Nucleic Acid Analogues ........................................................................ 20
Example 3(k)-1: Nucleotide ambiguity symbols .............................................................. 21
Example 3(k)-2: Ambiguity symbol "n" used in both a conventional and nonconventional manner ................................................................. 22
Example 3(k)-3: Ambiguity symbol "n" used in a nonconventional manner ..................... 23
Example 3(k)-4: Ambiguity symbols other than "n" are “specifically defined” ..................... 24
Example 11(a)-1: Double-stranded nucleotide sequence – same lengths ....................... 37
Example 11(b)-1: Double-stranded nucleotide sequence – different lengths .................. 38
Example 11(b)-2: Double-stranded nucleotide sequence – no base-pairing segment ....... 39
Example 14-1: The symbol "t" represents uracil in RNA ..................................................... 40
Example 87-1: Encoding nucleotide sequence and encoded amino acid sequence .......... 54
Example 91-1: Representation of enumerated variants .................................................... 57
Example 93(b)-1: Representation of individual variant sequences with multiple interdependent variations ................................................................. 63

Paragraph 7(b) – Amino acid sequences required in a sequence listing

Example 7(b)-1: Four or more specifically defined amino acids ....................................... 33
Example 7(b)-2: Branched amino acid sequence ............................................................. 34
Example 7(b)-3: Branched amino acid sequence ............................................................. 36

Cross-referenced examples
Example 3(a)-1: D amino acids ......................................................................................... 14
Example 3(c)-1: Enumeration of amino acids by chemical structure ............................... 15
Example 3(c)-2: Shorthand formula for an amino acid sequence .................................. 16
Example 3(k)-5: Ambiguity abbreviation “Xaa” used in a nonconventional manner ........................................ 25
Example 27-1: Shorthand formula for a nucleotide sequence ........................................................................ 42
Example 27-3: Shorthand formula - four or more specifically defined amino acids ........................................ 44
Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid ................................................ 47
Example 30-1: Feature key “CARBODHYD” .................................................................................................. 48
Example 36-1: Sequence with a region of a known number of “X” residues represented as a single sequence .......................................................................................................................... 49
Example 37-1: Sequence with regions of an unknown number of “X” residues must not be represented as a single sequence .................................................................................................................. 52
Example 37-2: Sequence with regions of an unknown number of “X” residues must not be represented as a single sequence .................................................................................................................. 53
Example 37-3: Sequence with regions of an unknown number of “X” residues must not be represented as a single sequence .................................................................................................................. 54
Example 91-2: Representation of enumerated variants ...................................................................................... 58
Example 91-3: Representation of a consensus sequence ...................................................................................... 59
Example 92-1: Representation of single sequence with enumerated alternative amino acids .................................. 61
Example 93(a)-1: Representation of a variant sequence by annotation of the primary sequence .......................... 62

Paragraph 8 – Threshold for inclusion of sequences

Cross-referenced examples

Example 3(k)-1: Nucleotide ambiguity symbols ................................................................................................. 21
Example 3(k)-2: Ambiguity symbol “n” used in both a conventional and nonconventional manner ................... 22
Example 7(a)-1: Branched nucleotide sequence .................................................................................................. 26
Example 7(a)-6: Nonconventional nucleotide symbols .......................................................................................... 32
Example 7(b)-1: Four or more specifically defined amino acids ........................................................................ 33
Example 14-1: The symbol “t” represents uracil in RNA ...................................................................................... 40
Example 37-1: Sequence with regions of an unknown number of “X” residues must not be represented as a single sequence .................................................................................................................. 52
Example 37-2: Sequence with regions of an unknown number of “X” residues must not be represented as a single sequence .................................................................................................................. 53
Example 92-1: Representation of single sequence with enumerated alternative amino acids .......................... 61

Paragraph 11 – Representation of a nucleotide sequence

Cross-referenced examples

Example 3(g)-4: Nucleic Acid Analogues ........................................................................................................... 20
Example 7(a)-1: Branched nucleotide sequence .................................................................................................. 26

Paragraph 11(a) – Double-stranded nucleotide sequence - fully complementary

Example 11(a)-1: Double-stranded nucleotide sequence – same lengths.......................................................... 37
Paragraph 11(b) – Double-stranded nucleotide sequence – not fully complementary

Example 11(b)-1: Double-stranded nucleotide sequence – different lengths ........................................... 38
Example 11(b)-2: Double-stranded nucleotide sequence – no base-pairing segment ................................ 39

Paragraph 13 – Representation of nucleotides

Cross-referenced examples

Example 3(k)-2: Ambiguity symbol “n” used in both a conventional and nonconventional manner .......... 22
Example 7(a)-1: Branched nucleotide sequence ......................................................................................... 26
Example 14-1: The symbol “t” represents uracil in RNA ......................................................................... 40
Example 91-1: Representation of enumerated variants .............................................................................. 57

Paragraph 14 – Symbol “t” construed as uracil in RNA

Example 14-1: The symbol “t” represents uracil in RNA ......................................................................... 40

Paragraph 15 – The most restrictive nucleotide ambiguity symbol should be used

Cross-referenced examples

Example 3(g)-1: Nucleotide sequence interrupted by a C3 spacer .............................................................. 17
Example 3(g)-2: Nucleotide sequence with residue alternatives, including a C3 spacer ......................... 18
Example 3(k)-4: Ambiguity symbols other than “n” are “specifically defined” .............................................. 24
Example 93(b)-1: Representation of individual variant sequences with multiple interdependent variations 63

Paragraph 16 – Representation of a modified amino acid

Cross-referenced examples

Example 3(g)-1: Nucleotide sequence interrupted by a C3 spacer .............................................................. 17
Example 3(g)-4: Nucleic Acid Analogue ................................. 20

Paragraph 17 – Annotation of a modified amino acid

Cross-referenced examples

Example 3(g)-1: Nucleotide sequence interrupted by a C3 spacer .............................................................. 17
Example 3(g)-3: Abasic site ......................................................................................................................... 19
Example 7(a)-1: Branched nucleotide sequence ......................................................................................... 26
Example 7(a)-6: Nonconventional nucleotide symbols ............................................................................ 32

Paragraph 18 – Annotation of regions of consecutive modified nucleotides

Cross-referenced examples

Example 3(g)-4: Nucleic Acid Analogue ................................................................. 20
Example 11(b)-1: Double-stranded nucleotide sequence – different lengths ........................................ 39
Paragraph 19 – Annotation of uracil in DNA or thymine in RNA

Cross-referenced examples

Example 14-1: The symbol “t” represents uracil in RNA

Paragraph 25 – Amino acid sequence residue position number 1

Cross-referenced examples

Example 3(a)-1: D amino acids
Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid

Paragraph 26 – Representation of amino acids

Cross-referenced examples

Example 36-1: Sequence with a region of a known number of “X” residues represented as a single sequence
Example 87-1: Encoding nucleotide sequence and encoded amino acid sequence
Example 90-1: Amino acid sequence encoded by a coding sequence with introns
Example 91-2: Representation of enumerated variants
Example 91-3: Representation of a consensus sequence

Paragraph 27 – The most restrictive amino acid ambiguity symbol should be used

Example 27-1: Shorthand formula for a nucleotide sequence
Example 27-2: Shorthand formula - less than four specifically defined amino acids
Example 27-3: Shorthand formula - four or more specifically defined amino acids

Cross-referenced examples

Example 7(b)-1: Four or more specifically defined amino acids
Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid
Example 36-1: Sequence with a region of a known number of “X” residues represented as a single sequence
Example 36-2: Sequence with multiple regions of a known number or range of “X” residues represented as a single sequence
Example 36-3: Sequence with multiple regions of a known number or range of “X” residues represented as a single sequence
Example 91-3: Representation of a consensus sequence
Example 92-1: Representation of single sequence with enumerated alternative amino acids
Example 93(a)-1: Representation of a variant sequence by annotation of the primary sequence

Paragraph 28 – Amino acid sequences separated by internal terminator symbols

Example 28-1: Encoding nucleotide sequence and encoded amino acid sequence

Cross-referenced examples

Example 87-1: Encoding nucleotide sequence and encoded amino acid sequence
Example 90-1: Amino acid sequence encoded by a coding sequence with introns ........................................ 55

**Paragraph 29 – Representation of an “other” modified amino acid**

**Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid** ........................................ 47

**Cross-referenced examples**

Example 3(a)-1: D amino acids ........................................................................................................ 14

Example 30-1: Feature key “CARBODHYD” ....................................................................................... 48

**Paragraph 30 – Annotation of a modified amino acids**

**Example 30-1: Feature key “CARBODHYD”** ................................................................................... 48

**Cross-referenced examples**

Example 3(a)-1: D amino acids ........................................................................................................ 14

Example 3(c)-1: Enumeration of amino acids by chemical structure .................................................... 15

Example 7(b)-2: Branched amino acid sequence .................................................................................. 34

Example 7(b)-3: Branched amino acid sequence .................................................................................. 36

**Paragraph 31 – Representation of a D-amino acid**

**Cross-referenced examples**

Example 3(a)-1: D amino acids ........................................................................................................ 14

Example 3(c)-1: Enumeration of amino acids by chemical structure .................................................... 15

Example 7(b)-2: Branched amino acid sequence .................................................................................. 34

Example 7(b)-3: Branched amino acid sequence .................................................................................. 36

**Paragraph 32 – Annotation of an “unknown” amino acid**

**Cross-referenced examples**

Example 3(c)-1: Enumeration of amino acids by chemical structure .................................................... 15

**Paragraph 34 – Annotation of a contiguous region of “X” residues**

**Cross-referenced examples**

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid ........................................ 47

**Paragraph 36 – Sequences containing regions of an exact number of contiguous “n” or “X” residues**

**Example 36-1: Sequence with a region of a known number of “X” residues represented as a single sequence** .......................................................... 49

**Example 36-2: Sequence with multiple regions of a known number or range of “X” residues represented as a single sequence** ............................................................................. 50

**Example 36-3: Sequence with multiple regions of a known number or range of “X” residues represented as a single sequence** ............................................................................. 51

**Paragraph 37 – Sequences containing regions of an unknown number of contiguous “n” or “X” residues**

**Example 37-1: Sequence with regions of an unknown number of “X” residues must not be represented as a single sequence** ............................................................................. 52
Example 37-2: Sequence with regions of an unknown number of “X” residues must not be represented as a single sequence ................................................................. 53

Paragraph 54 – The element INSDSeq_moltype

Cross-referenced examples
Example 14-1: The symbol “t” represents uracil in RNA ................................................................. 40

Paragraph 57 – The element INSDSeq_sequence

Cross-referenced examples
Example 28-1: Encoding nucleotide sequence and encoded amino acid sequence ................................. 45
Example 90-1: Amino acid sequence encoded by a coding sequence with introns .................................. 55

Paragraph 65 – Location descriptor

Cross-referenced examples
Example 3(g)-4: Nucleic Acid Analogues ......................................................................................... 20

Paragraph 66 – Location descriptor syntax

Cross-referenced examples
Example 3(g)-4: Nucleic Acid Analogues ......................................................................................... 20
Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid ..................................... 47

Paragraph 70 – Feature locations

Cross-referenced examples
Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid ..................................... 47

Paragraph 71 – Representation of the characters “<” and “>” in a location descriptor

Cross-referenced examples
Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid ..................................... 47

Paragraph 87 – “CDS” Feature key

Example 87-1: Encoding nucleotide sequence and encoded amino acid sequence ................................. 54

Cross-referenced examples
Example 90-1: Amino acid sequence encoded by a coding sequence with introns .................................. 55

Paragraph 88 – The qualifiers “transl_table” and “translation”

Cross-referenced examples
Example 28-1: Encoding nucleotide sequence and encoded amino acid sequence ................................. 45
Example 87-1: Encoding nucleotide sequence and encoded amino acid sequence ................................. 54
Example 90-1: Amino acid sequence encoded by a coding sequence with introns .................................. 55

Paragraph 90 – Encoded amino acid sequence inclusion in a sequence listing

Example 90-1: Amino acid sequence encoded by a coding sequence with introns .................................. 55
Cross-referenced examples

Example 28-1: Encoding nucleotide sequence and encoded amino acid sequence ........................................ 45

Example 87-1: Encoding nucleotide sequence and encoded amino acid sequence ......................................... 54

Paragraph 91 – Primary sequence and a variant each enumerated by its residue

Example 91-1: Representation of enumerated variants ....................................................................................... 57

Example 91-2: Representation of enumerated variants ....................................................................................... 58

Example 91-3: Representation of a consensus sequence ....................................................................................... 59

Paragraph 92 – Variant sequence disclosed as a single sequence with enumerated alternative residues

Example 92-1: Representation of single sequence with enumerated alternative amino acids ............................. 61

Paragraph 93(a) – A variant sequence disclosed only by reference to a primary sequence with multiple independent variations

Example 93(a)-1: Representation of a variant sequence by annotation of the primary sequence ....................... 62

Paragraph 93(b) – A variant sequence disclosed only by reference to a primary sequence with multiple interdependent variations

Example 93(b)-1: Representation of individual variant sequences with multiple interdependent variations ........ 63

Paragraph 94 – Feature keys and qualifiers for a variant sequence

Cross-referenced examples

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid ................................................. 47

Paragraph 95 – Annotation of a variant sequence

Cross-referenced examples

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid ................................................. 47

Example 91-3: Representation of a consensus sequence ....................................................................................... 59

Example 92-1: Representation of single sequence with enumerated alternative amino acids ............................. 60
EXAMPLES

Paragraph 3(a) Definition of “amino acid”

Example 3(a)-1: D amino acids

A patent application describes the following sequence:

Cyclo (D-Ala-D-Glu-Lys-Nle-Gly-D-Met-D-Nle)

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

Paragraph 3(a) of the Standard defines “amino acid” as including “D-amino acids” and amino acids containing modified or synthetic side chains. Based on this definition, the enumerated peptide contains five amino acids that are specifically defined (D-Ala, D-Glu, Lys, Gly, and D-Met). Therefore, the sequence must be included in a sequence listing as required by ST.26 paragraph 7(b).

Question 3: How should the sequence(s) be represented in the sequence listing?

Paragraph 29 requires that D-amino acids should be represented in the sequence as the corresponding unmodified L-amino acid. Further, any modified amino acid that cannot be represented by any other symbol in Annex I, Section 3, Table 3, must be represented by the symbol “X”.

In this example, the sequence contains three D-amino acids that can be represented by an unmodified L-amino acid in Annex I, Section 3, Table 3, one L-amino acid (Nle), and one D-amino acid (D-Nle) that must be represented by the symbol “X”.

Paragraph 25 indicates that when amino acid sequences are circular in configuration, applicant must choose the amino acid in residue position number 1. Accordingly, the sequence may be represented as:

AEKXGMX (SEQ ID NO: 1)

or otherwise, with any other amino acid in the sequence in residue position number 1. A feature key “SITE” and a qualifier “NOTE” must be provided for each D-amino acid with the complete, unabbreviated name of the D-amino acid as the qualifier value, e.g., D-Alanine and D-Norleucine. Further, a feature key “SITE” and a qualifier “NOTE” must be provided with the abbreviation for L-norleucine as the qualifier value, i.e. “Nle”, as set forth in Annex I, Section 4, Table 4. Finally, a feature key “REGION” and a qualifier “NOTE” should be provided to indicate that the peptide is circular.

Relevant ST.26 paragraphs: Paragraphs 3(a), 7(b), 25, 26, 29, 30, and 31
**Paragraph 3(c) – Definition of “enumeration of its residues”**

**Example 3(c)-1: Enumeration of amino acids by chemical structure**

![Chemical structure](image)

**Question 1: Does ST.26 require inclusion of the sequence(s)?**

**YES**

The enumerated peptide, illustrated as a structure, contains at least four specifically defined amino acids. Therefore, the sequence must be included in a sequence listing.

**Question 3: How should the sequence(s) be represented in the sequence listing?**

The sequence may be represented as:

**VAFXGK (SEQ ID NO: 2)**

wherein “X” represents an “other” modified amino acid: , which requires a feature key “SITE” together with the qualifier “NOTE”. The qualifier “NOTE” provides the complete, unabbreviated name of the modified tryptophan in position 4 of the enumerated peptide, e.g., “6-amino-7-(1H-indol-3-yl)-5-oxoheptanoic acid”. Further, additional feature keys “SITE” and qualifier “NOTE” are required to indicate the acetylation of the N-terminus and the methylation of the C-terminus.

Alternatively, the sequence may be represented as:

**VAFW (SEQ ID NO: 3)**

A feature key “SITE” and qualifier “NOTE” are required to indicate modification of tryptophan in position 4 of the enumerated peptide with the value: “C-terminus linked via a glutaraldehyde bridge to dipeptide GK”. Further, an additional feature key “SITE” at location 1 and qualifier “NOTE” is required to indicate the acetylation of the N-terminus.

**Relevant ST.26 paragraph(s): Paragraphs 3(c), 7(b), 29, 30, and 31**
Example 3(c)-2: Shorthand formula for an amino acid sequence

\[(G_4z)_n\]

Where G = Glycine, z = any amino acid and variable n can be any whole integer.

**Question 1:** Does ST.26 require inclusion of the sequence(s)?

**Yes**

The disclosure indicates that "n" can be "any whole integer"; therefore, the most encompassing embodiment of "n" is indeterminate. Since "n" is indeterminate, the peptide of the formula cannot be expanded to a definite length, and therefore, the unexpanded formula must be considered.

The enumerated peptide in the unexpanded formula ("n" = 1) provides four specifically defined amino acids, each of which is Gly, and the symbol "z". Conventionally, "z" is the symbol for "glutamine or glutamic acid", however, the example defines "z" as "any amino acid". Under ST.26, an amino acid that is not specifically defined is represented by "X". Based on this analysis, the enumerated peptide, i.e. GGGGX, contains four glycine residues that are enumerated and specifically defined. Thus, ST.26 paragraph 7(b) requires inclusion of the sequence in a sequence listing.

**Question 3:** How should the sequence(s) be represented in the sequence listing?

The sequence uses a nonconventional symbol "z", the definition of which must be determined from the disclosure (see Introduction to this document). Since "z" is defined as any amino acid, the conventional symbol used to represent this amino acid is "X". Therefore, the sequence must be represented as a single sequence:

GGGGX (SEQ ID NO: 4)

preferably annotated with the feature key REGION, feature location ">5" (corresponds to >5), with a NOTE qualifier with the value "The entire sequence of amino acids 1-5 can be repeated one or more times."

**CAUTION:** The preferred representation of the sequence indicated above is directed to the provision of a sequence listing on the filing date of a patent application. The same representation may not be applicable to a sequence listing provided subsequent to the filing date of a patent application, since consideration must be given to whether the information provided could be considered by an IPO to add subject matter to the original disclosure.

Relevant ST.26 paragraph(s): Paragraph 3(c) and 7(b)
Paragraph 3(g) Definition of “nucleotide”

Example 3(g)-1: Nucleotide sequence interrupted by a C3 spacer

A patent application describes the following sequence:

\[ \text{atgcatgcatgcatgc} \]

where \( n = \) a C3 spacer with the following structure:

![C3 Spacer Structure]

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

The enumerated sequence contains two segments of specifically defined nucleotides separated by a C3 spacer. The C3 spacer is not a nucleotide according to paragraph 3(g); the conventional symbol “\( n \)” is being used in a nonconventional manner (see Introduction to this document). Consequently, each segment is a separate nucleotide sequence. Since each segment contains more than 10 specifically defined nucleotides, both must be included in a sequence listing.

Question 3: How should the sequence(s) be represented in the sequence listing?

Each segment must be included in a sequence listing as a separate sequence, each with their own sequence identification number:

- atgcatgcatgcatgc (SEQ ID NO: 5)
- cggcatgcatgcatgc (SEQ ID NO: 6)

The cytosine in each segment that is attached to the C3 spacer should be further described in a feature table using the feature key “misc_feature” and the qualifier “note”. The “note” qualifier value, which is “free text”, should indicate the presence of the spacer, which is joined to another nucleic acid.

Relevant ST.26 paragraphs: Paragraphs 3(g), 7(a), and 15
Example 3(g)-2: Nucleotide sequence with residue alternatives, including a C3 spacer

A patent application describes the following sequence:

atgcatgcatgcncggcatgcatgc

where \( n = c, a, g, \) or a C3 spacer with the following structure:

![C3 Spacer Diagram](image)

Question 1: Does ST.26 require inclusion of the sequence(s)?

**YES**

There are 24 specifically defined residues in the enumerated sequence interrupted by the variable “\( n \).” The explanation of the sequence in the disclosure must be consulted to determine if the “\( n \)” is used in a conventional or nonconventional manner (see Introduction to this document).

The disclosure indicates that \( n = c, a, g, \) or a C3 spacer. The “\( n \)” is a conventional symbol used in a nonconventional manner, since it is described as including a C3 spacer, which does not meet the definition of a nucleotide. The symbol “\( n \)” is also described as including “c,” “a,” or “g”; therefore, ST.26 requires inclusion of the 25 nucleotide sequence in a sequence listing. Since two segments separated by the C3 spacer are distinct sequences from the 25 nucleotide sequence, the two 12 nucleotide sequences may also be included.

Question 3: How should the sequence(s) be represented in the sequence listing?

The example indicates that “\( n = c, a, g, \) or a C3 spacer”. As discussed above, a C3 spacer is not a nucleotide. According to paragraph 15, the symbol “\( n \)” must not be used to represent anything other than a nucleotide; therefore, the symbol “\( n \)” cannot represent a C3 spacer in a sequence listing.

Paragraph 15 also states that where an ambiguity symbol is appropriate, the most restrictive symbol should be used. The symbol “\( v \)” represents “a or c or g” according to Annex I, Section 1, Table 1, which is more restrictive than “\( n \)”.

Where variable “\( n \)” in the example is c, a, or g, the single sequence enumerated by its residues that includes the most disclosed embodiments, and is therefore, the most encompassing sequence (see Introduction to this document) that must be included in a sequence listing is:

atgcatgcatgcvcggcatgcatgc (SEQ ID NO: 7)

Inclusion of any additional sequences essential to the disclosure or claims of the invention is strongly encouraged, as discussed in the introduction to this document.

Where variable “\( n \)” in the example is a C3 spacer, the sequence can be considered two separate segments of specifically defined nucleotides on either side of the variable “\( n \)”, i.e., atgcatgcatgc (SEQ ID NO: 8); and cggcatgcatgc (SEQ ID NO: 9). If essential to the disclosure or claims, these two sequences should also be included in the sequence listing, each with their own sequence identification number.

The cytosine in each segment that is attached to the C3 spacer should be further described in a feature table using the feature key “misc_feature” and the qualifier “note”. The “note” qualifier value, which is “free text”, should indicate the presence of the spacer, which is joined to another nucleic acid and identify the spacer by either its complete unabbreviated chemical name, or by its common name, e.g., C3 spacer.

CAUTION: The preferred representation of the sequence indicated above is directed to the provision of a sequence listing on the filing date of a patent application. The same representation may not be applicable to a sequence listing provided subsequent to the filing date of a patent application, since consideration must be given to whether the information provided could be considered by an IPO to add subject matter to the original disclosure.

Relevant ST.26 paragraphs: Paragraphs 3(g), 7(a), and 15
Example 3(g)-3: Abasic site

A patent application describes the following sequence:

gagcattgac-AP-taaggct

Wherein AP is an abasic site

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

The specifically defined residues of the enumerated sequence are interrupted by an abasic site. The 5' side of the abasic site contains 10 nucleotides and the 3' side of the abasic site contains 7 nucleotides. Paragraph 3(g)(ii)(2) defines an abasic site as a "nucleotide" when it is part of a nucleotide sequence. Consequently, the abasic site in this example is considered a "nucleotide" for the purposes of determining if and how the sequence is required to be included in a sequence listing. Accordingly, the residues on each side of the abasic site are part of a single enumerated sequence containing 18 nucleotides total, 17 of which are specifically defined. Therefore, the sequence must be included as a single sequence in a sequence listing as required by ST.26 paragraph (7)(b).

Question 3: How should the sequence(s) be represented in the sequence listing?

The sequence must be included in a sequence listing as:

```
gagcattgacntaaggct (SEQ ID NO: 10)
```

The abasic site must be represented by an "n" and must be further described in a feature table. The preferred means of annotation is the feature key "modified_base" and the mandatory qualifier "mod_base" with the value "OTHER". A "note" qualifier must be included that describes the modified base as an abasic site.

Relevant ST.26 paragraphs: Paragraphs 3(g), 7(a), and 17
Example 3(g)-4: Nucleic Acid Analogues

A patent application discloses the following glycol nucleic acid (GNA) sequence:

PO₄-tagttcattqactaaggctccccattqact-OH

Wherein the left end of the sequence mimics the 5’ end of a DNA sequence.

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES – The individual residues that comprise a GNA sequence are considered nucleotides according to ST.26 paragraph 3(g)(i)(2). Accordingly, the sequence has more than ten enumerated and “specifically defined” nucleotides and is required to be included in a sequence listing.

Question 3: How should the sequence(s) be represented in the sequence listing?

GNA sequences do not have a 5’-end and a 3’-end, but rather, a 3’-end and a 2’-end. The 3’-end, which is routinely depicted as having a terminal phosphate group, corresponds to the 5’-end of DNA or RNA. (Note that other nucleic acid analogues may correspond differently to the 5’-end and 3’-end of DNA and RNA.) According to paragraph 10, it must be included in a sequence listing “in the direction from left to right that mimics the 5’-end to 3’-end direction.” Therefore, it must be included in a sequence listing as:

tagttcattgactaaggctccccattgact (SEQ ID NO: 11)

The sequence must be described in a feature table using the feature key “modified_base” and the mandatory qualifier “mod_base” with the abbreviation “OTHER”. A “note” qualifier must be included with the complete unabbreviated name of the modified nucleotides, such as “glycol nucleic acids” or “2,3-dihydroxypropyl nucleosides”. A single INSDFeature element can be used to describe the entire sequence as a GNA where the INSDFeature_location has the range “1..30”.

Relevant ST.26 paragraphs: Paragraphs 3(d), 3(g), 7(a), 11, 16, 18, 65, and 66
Paragraph 3(k) Definition of “specifically defined”

Example 3(k)-1: Nucleotide ambiguity symbols

5' NNG KNG KNG K 3'

N and K are IUPAC-IUB ambiguity codes

Question 1: Does ST.26 require inclusion of the sequence(s)?

NO

IUPAC-IUB ambiguity codes correspond to the list of nucleotide symbols defined in Annex I, Section 1, Table 1. According to paragraph 3(k), a specifically defined nucleotide is any nucleotide other than those represented by the symbol “n” listed in Annex I. Therefore, “K” and “G” are specifically defined nucleotides and “N” is not a specifically defined nucleotide.

The enumerated sequence does not have ten or more specifically defined nucleotides and therefore is not required by ST.26 paragraph 7(a) to be included in a sequence listing.

Question 2: Does ST.26 permit inclusion of the sequence(s)?

NO

According to paragraph 8, “A sequence listing must not include any sequences having fewer than ten specifically defined nucleotides…” The enumerated sequence does not have ten or more specifically defined nucleotides; therefore, it must not be included in a sequence listing.

Relevant ST.26 paragraphs: Paragraphs 3(k), 7(a), 8, and 13
Example 3(k)-2: Ambiguity symbol “n” used in both a conventional and nonconventional manner

An application discloses the artificial sequence: 5'-AATGCCGGAN-3'. The disclosure further states:

(i) in one embodiment, N is any nucleotide;
(ii) in one embodiment, N is optional but is preferably G;
(iii) in one embodiment, N is K;
(iv) in one embodiment, N is C.

Question 1: Does ST.26 require inclusion of the sequence(s)?

NO

The enumerated sequence contains 9 specifically defined nucleotides and an "N." The explanation of the sequence in the disclosure must be consulted to determine if the symbol "N" is used in a conventional manner (see Introduction to this document).

Consideration of disclosed embodiments (i) through (iv) of the enumerated sequence reveals that the most encompassing embodiment of "N" is "any nucleotide." In the most encompassing embodiment, "N" in the enumerated sequence is used in a conventional manner.

In certain embodiments "N" is described as specifically defined residues (i.e., "N is C" in part (iv)). However, only the most encompassing embodiment (i.e., "N is any nucleotide") is considered when determining if a sequence must be included in a sequence listing. Thus, the enumerated sequence that must be evaluated is 5'-AATGCCGGAN-3'.

Based on this analysis, the enumerated sequence, i.e. AATGCCGGAN, does not contain ten specifically defined nucleotides. Therefore, ST.26 paragraph 7(a) does not require inclusion of the sequence in a sequence listing, despite the fact that "n" is also defined as specific nucleotides in some embodiments.

Question 2: Does ST.26 permit inclusion of the sequence(s)?

NO

The sequence “AATGCCGGAN” must not be included in a sequence listing.

However, a described alternative sequence may be included in a sequence listing if the “N” is replaced with a specifically defined nucleotide.

Question 3: How should the sequence(s) be represented in the sequence listing?

Inclusion of sequences which represent embodiments that are a key part of the invention is strongly encouraged. Inclusion of these sequences allows for a more thorough search and provides public notice of the subject matter for which a patent is sought.

For the above example, it is highly recommended that the following three additional sequences are included in the sequence listing, each with their own sequence identification number:

aatgccgga (SEQ ID NO: 12)
aatgccgga (SEQ ID NO: 13)
aatgccgca (SEQ ID NO: 14)

If less than all three of the above sequences are included, the nucleotide that replaces the “n” should be annotated to describe the alternatives. For example, if only SEQ ID NO: 12 above is included in the sequence listing, the feature key “misc_difference” with feature location “10” should be used together with two “replace” qualifiers where the value for one would be “g” and the second would be “c”.

CAUTION: The preferred representation of the sequence indicated above is directed to the provision of a sequence listing on the filing date of a patent application. The same representation may not be applicable to a sequence listing provided subsequent to the filing date of a patent application, since consideration must be given to whether the information provided could be considered by an IPO to add subject matter to the original disclosure.

Relevant ST.26 paragraphs: Paragraphs 3(k), 7(a), 8, and 13
Example 3(k)-3: Ambiguity symbol "n" used in a nonconventional manner

An application discloses the sequence: 5'-aatgtggn-3'

Wherein n is c

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

According to paragraph 3(k), a “specifically defined” nucleotide is any nucleotide other than those represented by the symbol “n” listed in Annex I, Section 1, Table 1.

In this example “n” is used in a nonconventional manner to represent only “c”. The disclosure does not indicate that “n” is used in the conventional manner to represent “any nucleotide”. Therefore, the sequence must be interpreted as if the equivalent conventional symbol, i.e. "c", had been used in the sequence (see Introduction to this document). Accordingly, the enumerated sequence that must be considered is:

5'-aatgtgga-c-3'

This sequence has ten specifically defined nucleotides and is required by ST.26 paragraph 7(a) to be included in a sequence listing.

Question 3: How should the sequence(s) be represented in the sequence listing?

The sequence must be included in a sequence listing as: aatgtgga-c (SEQ ID NO: 15)

Relevant ST.26 paragraphs: Paragraphs 3(k) and 7(a)
Example 3(k)-4: Ambiguity symbols other than “n” are “specifically defined”

A patent application describes the following sequence:

5' NNG KNG KNG KAG VCR 3'

wherein N, K, V, and R are IUPAC-IUB ambiguity codes

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

IUPAC-IUB ambiguity codes correspond to the list of nucleotide symbols defined in Annex I, Section 1, Table 1. According to paragraph 3(k), a “specifically defined” nucleotide is any nucleotide other than those represented by the symbol “n” listed in Annex I, Section 1, Table 1. Therefore, “K”, “V”, and “R” are “specifically defined” nucleotides.

The sequence has eleven enumerated and “specifically defined” nucleotides and is required by ST.26 paragraph 7(a) to be included in a sequence listing.

Question 3: How should the sequence(s) be represented in the sequence listing?

The sequence must be included in a sequence listing as:

nngkngkngkagvcr (SEQ ID NO: 16)

Relevant ST.26 paragraphs: Paragraphs 3(k), 7(a) and 15
Example 3(k)-5: Ambiguity abbreviation “Xaa” used in a nonconventional manner

A patent application describes the following sequence:

Xaa-Tyr-Glu-Xaa-Xaa-Xaa-Leu

Wherein Xaa in position 1 is any amino acid, Xaa in position 4 is Lys, Xaa in position 5 is Gly and Xaa in position 6 is Leucine or Isoleucine.

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

The enumerated peptide in the formula provides three specifically defined amino acids in positions 2, 3 and 7. The first amino acid is represented by a conventional abbreviation, i.e., Xaa, representing any amino acid. However, the 4th, 5th and 6th amino acids are represented by a conventional abbreviation used in a nonconventional manner (see Introduction to this document). Therefore, the explanation of the sequence in the disclosure is consulted to determine the definition of “Xaa” in these positions. Since “Xaa” in positions 4-6 are indicated as a specific amino acid, the sequence must be interpreted as if the equivalent conventional abbreviations had been used in the sequence, i.e., Lys, Gly, and (Leu or Ile). Consequently, the sequence contains four or more specifically defined amino acids and must be included in a sequence listing as required by ST.26 paragraph 7(b).

Question 3: How should the sequence(s) be represented in the sequence listing?

The sequence uses a conventional abbreviation “Xaa” in a nonconventional manner. Therefore, the explanation of the sequence in the disclosure must be consulted to determine the definition of “Xaa” in positions 4, 5 and 6. The explanation defines “Xaa” as a lysine in position 4, a glycine in position 5 and a leucine or isoleucine in position 6. The conventional symbols for these amino acids are K, G, and J respectively. Therefore, the sequence should be represented as in the sequence listing as:

XYEKJL (SEQ ID NO: 17)

Relevant ST.26 paragraphs: Paragraphs 3(k), 7(b), 26, and 27
**Paragraph 7(a) – Nucleotide sequences required in a sequence listing**

**Example 7(a)-1: Branched nucleotide sequence**

The description discloses the following branched nucleotide sequence:

3'-CA(pnp)CACACA(pnp)CACACA(pnp)CACACACA-(5') NH——C(=O)CH₂ 3' = Segment A

wherein "pnp" is a linkage or monomer containing an bromoacetylamino functionality;

3'-CA(pnp)CACACA(pnp)CACACA(pnp)CACACACA-(5') NH——C(=O)CH₂ 3' is segment A;

SP(O)=(O)CACACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA 3' is segments B, C, and D; and

SP(O)=(O)CACATAGGCATCTCCTAGTGCAGGAAGA 3' is segment E;

**Question 1: Does ST.26 require inclusion of the sequence(s)?**

**YES** – the four vertical segments B-E must be included in a sequence listing.

**NO** – the horizontal segment A must not be included in a sequence listing.

The above figure is an example of a "comb-type" branched nucleic acid sequence containing five linear segments: the horizontal segment A and the four vertical segments B-E.

According to paragraph 7(a), the linear portions of branched nucleotide sequences containing ten or more specifically defined nucleotides, wherein adjacent nucleotides are joined 3' to 5', must be included in a sequence listing.

The four vertical segments B-E each contain more than ten specifically defined nucleotides, wherein adjacent nucleotides are joined 3' to 5', and therefore each is required to be included in a sequence listing.
In horizontal segment A, the linear portions of the nucleotide sequence are linked by the non-nucleotide moiety "pnp" and each of these linked linear portions contains fewer than ten specifically defined nucleotides. Therefore, since no portion of segment A contains ten or more specifically defined nucleotides wherein adjacent nucleotides are joined 3' to 5', they are not required ST.26 paragraph 7(a) to be included in a sequence listing.

**Question 2: Does ST.26 permit inclusion of the sequence(s)?**

According to paragraph 8, "A sequence listing must not include any sequences having fewer than ten specifically defined nucleotides..."

No portion of Segment A contains ten or more specifically defined nucleotides wherein adjacent nucleotides are joined 3' to 5'; therefore, it must not be included in a sequence listing as a separate sequence with its own sequence identification number.

However, segments B, C, D, and E may be annotated to indicate that they are linked to segment A.

**Question 3: How should the sequence(s) be represented in the sequence listing?**

Segments B, C, and D are identical and must be included in a sequence listing as a single sequence:

```
cacacaaaaaaaaaaaaaaaaaaaaaaaaa
```

(SEQ ID NO: 18)

The first "c" in the sequence should be further described as a modified nucleotide using the feature key "misc_feature" and the qualifier "note" with the value e.g., "This sequence is one of four branches of a branched polynucleotide."

Segment E must be included in a sequence listing as a single sequence:

```
cacataggcatctcctagtgcaggaaga
```

(SEQ ID NO: 19)

The first "c" in the sequence should be further described as a modified nucleotide using the feature key "misc_feature" and the qualifier "note" with the value e.g., "This sequence is one of four branches of a branched polynucleotide."

**Relevant ST.26 paragraph(s):** Paragraphs 7(a), 8, 11, 13, and 17
Example 7(a)-2: Linear nucleotide sequence having a secondary structure

Wherein Ψ is pseudouridine.

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

The nucleotide sequence contains seventy-three enumerated and specifically defined nucleotides. Thus, the example has ten or more "specifically defined" nucleotides, and as required by ST.26 paragraph (7)(a), must be included in a sequence listing.

Question 3: How should the sequence(s) be represented in the sequence listing?

Consultation of the disclosure indicates that "Ψ" is equivalent to pseudouridine. The only conventional symbol that can be used to represent pseudouridine is "n"; therefore, the "Ψ" is a nonconventional symbol used to represent the conventional symbol "n" (see Introduction to this document). Accordingly, the sequence must be interpreted to have two "n" symbols in place of the two "Ψ" symbols.

The symbol "u" must not be used to represent uracil in an RNA molecule in the sequence listing. According to paragraph 14, the symbol "t" will be construed as uracil in RNA. The sequence must be included as:

```
GCAGGATTGTACGCAGCTGGGAGAGCGCCAGACTGATANCTGGAGTCTGTGTCGATCCACAGAATTGCACCA (SEQ ID NO: 20)
```

The value of the mandatory "mol_type" qualifier of the mandatory "source" feature key is "tRNA". Additional information may be provided with feature key "tRNA" and any appropriate qualifier(s).

The "n" residues must be further described in a feature table using the feature key "modified_base" and the mandatory qualifier "mod_base" with the abbreviation "p" for pseudouridine as the qualifier value (see Annex 1, Table 2).

Relevant ST.26 paragraph(s): Paragraphs 7(a), 11, 13, 14, 62, 84 and Annex I, sections 2 and 5, feature key 5.43
Example 7(a)-3: Nucleotide ambiguity symbols used in a nonconventional manner

A patent application describes the following sequence:

5’ GATC-MDR-MDR-MDR-MDR-GTAC 3’

The explanation of the sequence in the disclosure further indicates: “A “DR Element” consists of the sequence 5’ ATCAGCCAT 3’. A mutant DR Element, or MDR, is a DR element wherein the middle 5 nucleotides, CAGCC, are mutated to TTTTT.”

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

The enumerated sequence uses the symbol “MDR”. Where it is unclear if a symbol used in a sequence is intended to be a conventional symbol, i.e., a symbol set forth in Annex 1, Section 3, Table 3, or a nonconventional symbol, the explanation of the sequence in the disclosure must be consulted to make a determination (see Introduction to this document). According to Table 3, “MDR” could be interpreted as three conventional symbols (m = a or c, d = a or g or t/u, r = g or a) or as an abbreviation that is short-hand notation for some other structure.

Consultation of the disclosure indicates that an MDR element is equivalent to 5’ ATTTTTTAT 3’. The letters “MDR” are considered conventional symbols used in a nonconventional manner; therefore, the sequence must be interpreted as though it were disclosed using the equivalent conventional symbols. Accordingly, the enumerated sequence that is considered for inclusion in a sequence listing is:

5’ GATC ATTTTTAT ATTTTTAT ATTTTTAT ATTTTTAT GTAC 3’

The enumerated sequence has 44 specifically defined nucleotides and is required by ST.26 paragraph 7(a) to be included in a sequence listing.

Question 3: How should the sequence(s) be represented in the sequence listing?

The sequence must be included in a sequence listing as:

gatcattttttatattttttatattttttatattttttatgtac (SEQ ID NO: 21)

Relevant ST.26 paragraphs: Paragraph 7(a) and 13
Example 7(a)-4: Nucleotide ambiguity symbols used in a nonconventional manner

A patent application describes the following sequence:

5' ATTC-N-N-N-N-GTAC 3'

The explanation of the sequence in the disclosure further indicates that "N" consists of the sequence 5' ATACGCACT 3'.

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

The enumerated sequence uses the symbol "N". The explanation of the sequence in the disclosure must be consulted to determine if the "N" is used in a conventional or nonconventional manner (see Introduction to this document).

Consultation of the disclosure indicates that "N" is equivalent to 5' ATACGCACT 3'. Thus, the "N" is a conventional symbol used in a nonconventional manner. Accordingly, the sequence must be interpreted as though it were disclosed using the equivalent conventional symbols:

5' ATTC-ATACGCACT-ATACGCACT-ATACGCACT-ATACGCACT-GTAC 3'

The enumerated sequence has 44 specifically defined nucleotides and is required by ST.26 paragraph 7(a) to be included in a sequence listing.

Question 3: How should the sequence(s) be represented in the sequence listing?

The sequence must be included in a sequence listing as:

atctatgacactacgtacatacgtacatacgtacatacgtacatacgtacagtac (SEQ ID NO: 22)

Relevant ST.26 paragraphs: Paragraph 7(a) and 13
Example 7(a)-5: Nonconventional nucleotide symbols

A patent application describes the following sequence:

5' GATC-β-β-β-β-GTAC 3'

The explanation of the sequence in the disclosure further indicates that "β" consists of the sequence 5' ATACGCACT 3'.

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

The enumerated sequence uses the nonconventional symbol "β". The explanation of the sequence in the disclosure must be consulted to determine the meaning of "β" (see Introduction to this document).

Consultation of the disclosure indicates that "β" is equivalent to 5' ATACGCACT 3'. Thus, the "β" is a nonconventional symbol used to represent a sequence of nine specifically defined, conventional symbols. Accordingly, the sequence must be interpreted as though it were disclosed using the equivalent conventional symbols:

5' GATC-ATACGCACT-ATACGCACT-ATACGCACT-ATACGCACT-GTAC 3'

The enumerated sequence has 44 specifically defined nucleotides and is required by ST.26 paragraph 7(a) to be included in a sequence listing.

Question 3: How should the sequence(s) be represented in the sequence listing?

The sequence must be included in a sequence listing as:

gatcatacgcactatacgcactatacgcactatacgcactgtac (SEQ ID NO: 23)

Relevant ST.26 paragraphs: Paragraph 7(a) and 13
Example 7(a)-6: Nonconventional nucleotide symbols

A patent application describes the following sequence:

5' GATC-β-β-β-β-GTAC 3'

The explanation of the sequence in the disclosure further indicates that "β" is equal to adenine, inosine, or pseudouridine.

**Question 1:** Does ST.26 require inclusion of the sequence(s)?

**NO**

The enumerated sequence uses the nonconventional symbol "β". The explanation of the sequence in the disclosure must be consulted to determine the meaning of "β" (see Introduction to this document).

Consultation of the disclosure indicates that "β" is equivalent to adenine, inosine, or pseudouridine. The only conventional symbol that can be used to represent "adenine, inosine, or pseudouridine" is "n"; therefore, the "β" is a nonconventional symbol used to represent the conventional symbol "n". Accordingly, the sequence must be interpreted to have four "n" symbols in place of the four "β" symbols:

5' GATC-N-N-N-GTAC 3'

The enumerated sequence has only eight specifically defined nucleotides and is not required by ST.26 paragraph 7(a) to be included in a sequence listing.

**Question 2:** Does ST.26 permit inclusion of the sequence(s)?

**NO**

The enumerated sequence, 5' GATC-N-N-N-N-GTAC 3', must not be included in a sequence listing.

However, a disclosed alternative sequence may be included in a sequence listing if at least 2 of the "n" symbols are replaced by adenine, resulting in a sequence with at least 10 or more specifically defined nucleotides.

**Question 3:** How should the sequence(s) be represented in the sequence listing?

One possible permitted representation is:

gatcaaaagtac (SEQ ID NO: 24)

In the above example, the four adenine nucleotides that replace the β symbols should be annotated to note that these positions could be substituted with inosine or pseudouridine.

The feature key "misc_difference" should be used with a feature location 5-8 and a qualifier "note" with the value, e.g., "A nucleotide in any of positions 5-8 may be replaced with inosine or pseudouridine". Since these alternatives are modified nucleotides, then the feature key "modified_base" together with the qualifier "mod_base" would be required. The value for the "mod_base" qualifier can be "OTHER" with a "note" qualifier and the value of "i or p".

Other permutations are possible.

CAUTION: The preferred representation of the sequence indicated above is directed to the provision of a sequence listing on the filing date of a patent application. The same representation may not be applicable to a sequence listing provided subsequent to the filing date of a patent application, since consideration must be given to whether the information provided could be considered by an IPO to add subject matter to the original disclosure.

Relevant ST.26 paragraphs: Paragraph 7(a), 8, 13, and 17
Paragraph 7(b) – Amino Acid sequences required in a sequence listing

Example 7(b)-1: Four or more specifically defined amino acids

```
XXXXXXXXXDXXXXXXXXXXXXXFXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXGXXXXX
```

Where X = any amino acid

**Question 1:** Does ST.26 require inclusion of the sequence(s)?

**YES**

The enumerated peptide contains four specifically defined amino acids. The symbol "X" is used conventionally to represent the remaining amino acids as any amino acid (see Introduction to this document).

Because there are four specifically defined amino acids, i.e., Asp, Phe, Ala and Gly, ST.26 paragraph 7(b) requires that the sequence be included in a sequence listing.

**Question 3:** How should the sequence(s) be represented in the sequence listing?

The sequence must be represented as:

```
XXXXXXXXXDXXXXXXXXXXXXXFXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXGXXXXX
```

(SEQ ID NO: 25)

Since "X" can be any amino acid, annotation of the "X" residues is not required under paragraph 27.

**Relevant ST.26 paragraph(s):** Paragraphs 7(b), 8 and 27
Example 7(b)-2: Branched amino acid sequence

The application describes a branched sequence where the Lysine residues are used as a scaffolding core to form eight branches to which multiple linear peptide chains are attached. Lysine is a dibasic amino acid, providing it with two sites for peptide-bonding. The peptide is illustrated as follows:

In the above branched peptide, the bonds depicted by ———— represent an amide linkage between the terminal amine of the Lysine and the carboxyl end of the bonded amino acid. The bonds depicted by ^^^^^^^ represent an amide linkage between the side chain amine of the Lysine and the carboxyl end the bonded amino acid.

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

The example discloses a branched sequence where the lysine residues are used as a scaffolding. Paragraph 7(b) requires that the unbranched or linear portion of the sequence, containing four or more specifically defined amino acids, be included in a sequence listing. In the above example, the linear portions of the branched peptide that have four or more specifically defined amino acids are encircled.
ST.26 paragraph 7(b) requires inclusion of peptides 1-6 above in a sequence listing.

Peptides which are not required, and in fact are prohibited, from inclusion in the sequence listing are:

- YFA
- LLK

**Question 3: How should the sequence(s) be represented in the sequence listing?**

Peptides 1-6 must be represented with separate sequence identifiers:

- RISL (SEQ ID NO: 26)
- LLKK (SEQ ID NO: 27)
- IPACTA (SEQ ID NO: 28)
- FRAGGK (SEQ ID NO: 29)
- HQYFA (SEQ ID NO: 30)
- ATFGKKKA (SEQ ID NO: 31)

The cross linkage is preferably noted using the feature key "SITE" and the mandatory qualifier "NOTE" with the value e.g., "This sequence is one part of a branched peptide".

**Relevant ST.26 paragraph(s):** Paragraphs 7(b), 26, 30, and 31
Example 7(b)-3: Branched amino acid sequence

Peptide of the following sequence:

\[
\begin{align*}
\text{NH}_2\text{ASP--GLY--SER--ALA--LYS--LYS--LYS--LYS--CO}_2\text{H} \\
\text{NH}_2\text{ALAA--ALA--SER--HIS--GLY}
\end{align*}
\]

The linkage between the terminal Glycine residue in the lower sequence and the Lysine in the upper sequence is through an amide bond between the carboxy terminus of the Glycine and the amino terminal side chain of the Lysine.

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

The unbranched or linear portion of a sequence, containing four or more specifically defined amino acids, must be included in a sequence listing. In the above example, the linear portions of the branched peptide that have more than four amino acids are:

1. \[
\begin{align*}
\text{NH}_2\text{ASP--GLY--SER--ALA--LYS--LYS--LYS--LYS--CO}_2\text{H} \\
\text{NH}_2\text{ALAA--ALA--SER--HIS--GLY}
\end{align*}
\]

ST.26 paragraph 7(b) requires inclusion of peptides 1 and 2 in a sequence listing.

Question 3: How should the sequence(s) be represented in the sequence listing?

Peptides 1 and 2 must be represented with separate sequence identifiers:

DGSAKKK (SEQ ID NO: 32)

AASHG (SEQ ID NO: 33)

Preferably the sequence DGSAKKK should include an annotation to indicate that the 5th lysine is a modified amino acid using the feature key "SITE" together with the qualifier "NOTE" describing that lysine links the peptide AASHG. Preferably the sequence AASHG should include an annotation to indicate that the 5th glycine is linked to DGSAKKK using the feature key "SITE" together with the qualifier "NOTE".

Relevant ST.26 paragraph(s): Paragraphs 7(b), 26, 30, and 31
Paragraph 11(a) — Double-stranded nucleotide sequence — fully complementary

Example 11(a)-1: Double-stranded nucleotide sequence — same lengths

A patent application describes the following double-stranded DNA sequence:

\[
\begin{align*}
3' & \text{-} \text{CGCGTAAACGCTA} - 5' \\
5' & \text{-} \text{GGCCAATTGCAGAT} - 3'
\end{align*}
\]

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

Each enumerated nucleotide sequence has more than 10 specifically defined nucleotides. At least one strand must be included in the sequence listing, because the two strands of this double-stranded nucleotide sequence are fully complementary to each other.

Question 2: Does ST.26 permit inclusion of the sequence(s)?

While the sequence of only one strand must be included in the sequence listing, the sequences of both strands may be included, each with its own sequence identification number.

Question 3: How should the sequence(s) be represented in the sequence listing?

The double-stranded DNA sequence must be represented either as a single sequence or as two separate sequences. Each sequence included in the sequence listing must be represented in the 5' to 3' direction and assigned its own sequence identification number:

atgcaatggcc (top strand) (SEQ ID NO: 34)
and/or
ggccaattgcagat (bottom strand) (SEQ ID NO: 35)

Relevant ST.26 paragraphs: Paragraphs 7(a), 11(a), and 13
Paragraph 11(b) – Double-stranded nucleotide sequence - not fully complementary

Example 11(b)-1: Double-stranded nucleotide sequence – different lengths

A patent application contains the following drawing and caption:

```
5'-tagttcattgactaaggctccccattgactaaggcgactagcattgactaaggcaagc

```

The human gene ABC1 promoter region (top strand) bound by a PNA probe (bottom strand). Where “n” in the PNA probe is a universal PNA base selected from the group consisting of 5-nitroindole and 3-nitroindole.

**Question 1:** Does ST.26 require inclusion of the sequence(s)?

**YES** – the ABC1 promoter region (top strand)

The top strand has more than ten enumerated and “specifically defined” nucleotides and is required to be included in a sequence listing.

**YES** – the PNA probe (bottom strand)

The bottom strand must also be included in the sequence listing, with its own sequence identification number, because the two strands are not fully complementary to each other. The individual residues that comprise a PNA or “peptide nucleic acid” are considered nucleotides according to ST.26 paragraph 3(g). Therefore, the bottom strand has more than 10 enumerated and “specifically defined” nucleotides and is required to be included in a sequence listing.

**Question 3:** How should the sequence(s) be represented in the sequence listing?

The top strand must be included in a sequence listing as:

```
tagttcattgactaaggctccccattgactaaggcgactagcattgactaaggcaagc (SEQ ID NO: 36)
```

The bottom strand is a peptide nucleic acid and therefore does not have a 3' and 5' end. According to paragraph 11, it must be included in a sequence listing “in the direction from left to right that mimics the 5’-end to 3’-end direction.” Therefore, it must be included in a sequence listing as:

```
cgcctnagtcaatggg (SEQ ID NO: 37)
```

The “organism” qualifier of the feature key “source” must have the value “synthetic construct” and the mandatory qualifier “mol_type” with the value “other DNA”. The bottom strand must be described in a feature table using the feature key “modified_base” and the mandatory qualifier “mod_base” with the abbreviation “OTHER”. A “note” qualifier must be included with the complete unabbreviated name of the modified nucleotides, such as “N-(2-aminoethyl) glycine nucleosides”.

The “n” residue must be further described in a feature table using the feature key “modified_base” and the mandatory qualifier “mod_base” with the abbreviation “OTHER”. A “note” qualifier must be included with the complete unabbreviated name of the modified nucleotide: “N-(2-aminoethyl) glycine 5-nitroindole or N-(2-aminoethyl) glycine 3-nitroindole”.

**Relevant ST.26 paragraphs:** Paragraphs 3(g), 7(a), 11(b), 17, and 18
Example 11(b)-2: Double-stranded nucleotide sequence – no base-pairing segment

A patent application describes the following double-stranded DNA sequence:

3'−CCGGTTAGCTTATACGCTAGGGCTA−5'

5'−GGCCAATATGGCTTGCGATCCCGAT−3'

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

Each strand of the enumerated, double-stranded nucleotide sequence has more than 10 specifically defined nucleotides. Both strands must be included in the sequence listing, each with its own sequence identification number, because the two strands are not fully complementary to each other.

Question 3: How should the sequence(s) be represented in the sequence listing?

The sequence of each strand must be represented in the 5' to 3' direction and assigned its own sequence identification number:

atcgggatcgcatattcgattggcc (top strand) (SEQ ID NO: 38)

and

ggccaatatggcttgcgatcccgat (bottom strand) (SEQ ID NO: 39)

Relevant ST.26 paragraphs: Paragraphs 7(a), 11(b), and 13
Paragraph 14 – Symbol “t” construed as uracil in RNA

Example 14-1: The symbol “t” represents uracil in RNA

A patent application describes the following compound:

![Chemical structure diagram]

Wherein segment A and segment B are RNA sequences.

**Question 1:** Does ST.26 require inclusion of the sequence(s)?

- **YES – segment B**
- **NO – segment A**

The enumerated sequence contains two segments of specifically defined nucleotides separated by the following “linker” structure:

![Chemical structure diagram]

The linker structure is not a nucleotide according to paragraph 3(g); therefore, each segment must be considered a separate sequence. Segment B contains more than 10 specifically defined nucleotides and ST.26 paragraph 7(a) requires inclusion in a sequence listing. Segment A contains only 8 specifically defined nucleotides and therefore is not required to be included in a sequence listing.

**Question 2:** Does ST.26 permit inclusion of the sequence(s)?

- Segment A contains fewer than 10 specifically defined nucleotides, and therefore it must not be included in a sequence listing.
**Question 3: How should the sequence(s) be represented in the sequence listing?**

Segment B is an RNA molecule; therefore, the element “INSDSeq_moltype” must be “RNA.” The symbol “u” must not be used to represent uracil in an RNA molecule in a sequence listing. According to paragraph 14, the symbol “t” will be construed as uracil in RNA. Accordingly, segment B must be included in the sequence listing as:

```
tcctgctcggagatgttgat (SEQ ID NO: 40)
```

Thymine in RNA is considered a modified nucleotide, i.e. modified uracil, and must be represented in the sequence as “t” and be further described in a feature table. Accordingly, the thymine in position 1 must be further described using the feature key “modified_base”, the qualifier “mod_base” with “OTHER” as the qualifier value, and a qualifier “note” with “thymine” as the qualifier value.

```
The thymine, i.e. modified uracil, in position 1 should also be further described in a feature table using the feature key “misc_feature” and a qualifier “note” with the value e.g., “ccugucgt (Segment A) is attached at its 3’-end to a linker which is attached to the 5’ oxygen of the thymidine. The linker is (4-(3-hydroxybenzamido)butyl) phosphinic acid.”
```

**Relevant ST.26 paragraphs:** Paragraphs 3(g), 7(a), 8, 13, 14, 19, and 54
**Paragraph 27 – The most restrictive ambiguity symbol should be used**

**Example 27-1: Shorthand formula for a nucleotide sequence**

\[(\text{GGG}z)\]

Where \( z \) is any amino acid.

**Question 1: Does ST.26 require inclusion of the sequence(s)?**

**YES**

The sequence is disclosed as a formula. \((\text{GGG}z)\) is simply a shorthand way of representing the sequence \(\text{GGGzGGGz}\). Conventionally, a sequence is expanded first, and the definition of any variable, i.e. “\( z \)”, is determined thereafter.

The sequence uses the nonconventional symbol “\( z \)”. The definition of “\( z \)” must be determined from the explanation of the sequence in the disclosure, which defines this symbol as any amino acid (see Introduction to this document). The example does not provide any constraint on “\( z \)”, e.g., that it is the same in each occurrence.

Therefore, “\( z \)” is equivalent to the conventional symbol “\( X \)”, and the peptide in the example has eight enumerated amino acids, six of which are specifically defined glycine residues. ST.26 paragraph 7(b) requires inclusion of the sequence in a sequence listing as a single sequence with a single sequence identification number.

Note that the sequence is still encompassed by Paragraph 7(b) despite the fact that the enumerated and specifically defined residues are not contiguous.

**Question 3: How should the sequence(s) be represented in the sequence listing?**

The sequence uses the nonconventional symbol “\( z \)”, which according to the disclosure is any amino acid. The conventional symbol used to represent “any amino acid” is “\( X \)”. Therefore, the sequence must be represented as the single expanded sequence:

\[\text{GGGXGGGX} \text{ (SEQ ID NO: 41)}\]

Further, the example does not disclose that “\( z \)” is the same amino acid in both positions in the expanded sequence. However, if “\( z \)” is disclosed as the same amino acid in both positions, then a feature key “VARIANT” and a qualifier “NOTE” should be provided stating that “\( X \)” in position 4 and 8 can be any amino acid as long as they are the same in both positions.

**Relevant ST.26 paragraph(s):** Paragraphs 3(c), 7(b) and 27
Example 27-2: Shorthand formula - less than four specifically defined amino acids

A peptide of the formula (Gly-Gly-Gly-z)n.

The disclosure further states, that z is any amino acid and

(i) variable n is any length; or

(ii) variable n is 2-100, preferably 3

Question 1: Does ST.26 require inclusion of the sequence(s)?

NO

Consideration of both disclosed embodiments (i) and (ii) of the enumerated peptide of the formula reveals that “n” can be “any length”; therefore, the most encompassing embodiment of “n” is indeterminate. Since “n” is indeterminate, the peptide of the formula cannot be expanded to a definite length, and therefore, the unexpanded formula must be considered.

The enumerated peptide in the unexpanded formula (“n” = 1) provides three specifically defined amino acids, each of which is Gly, and the symbol “z”. Conventionally “Z” is the symbol for “glutamine or glutamic acid”; however, the example defines “z” as “any amino acid” (see Introduction to this document). Under ST.26, an amino acid that is not specifically defined is represented by “X”. Based on this analysis, the enumerated peptide, i.e. GGGX, does not contain four specifically defined amino acids. Therefore, ST.26 paragraph 7(b) does not require inclusion, despite the fact that “n” is also defined as specific numerical values in some embodiments.

Question 2: Does ST.26 permit inclusion of the sequence(s)?

YES

The example provides a specific numerical value for variable “n,” i.e., a lower limit of 2, an upper limit of 100, and an exact value 3. Any sequence containing at least four specifically defined amino acids may be included in the sequence listing.

Question 3: How should the sequence(s) be represented in the sequence listing?

A sequence containing 100 copies of GGGX is preferred (SEQ ID NO: 42). A further annotation should indicate that up to 98 copies of GGGX could be deleted. Inclusion of further specific embodiments that are a key part of the invention is strongly encouraged.

CAUTION: The preferred representation of the sequence indicated above is directed to the provision of a sequence listing on the filing date of a patent application. The same representation may not be applicable to a sequence listing provided subsequent to the filing date of a patent application, since consideration must be given to whether the information provided could be considered by an IPO to add subject matter to the original disclosure.

Relevant ST.26 paragraph(s): Paragraphs 3(c), 7(b), 26, and 27
Example 27-3: Shorthand formula - four or more specifically defined amino acids

A peptide of the formula (Gly-Gly-Gly-z)n,

Where z is any amino acid and variable n is 2-100, preferably 3.

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

The enumerated peptide of the formula provides three specifically defined amino acids, each of which is Gly, and the symbol “z.” Conventionally, “z” is the symbol for “glutamine or glutamic acid.” However, the description in this example defines “z” as “any amino acid” (see Introduction to this document). Under ST.26, an amino acid that is not specifically defined is represented by “X.” Based on this analysis, the enumerated repeat peptide does not contain four specifically defined amino acids. However, the description provides a specific numerical value for variable “n,” i.e., a lower limit of 2 and an upper limit of 100. Therefore, the example discloses a peptide having at least six specifically defined amino acids in the sequence GGGzGGGz, which is required by ST.26 to be included in a sequence listing.

Question 3: How should the sequence(s) be represented in the sequence listing?

Since “z” represents any amino acid, the conventional symbol used to represent the fourth and eighth amino acids is “X.”

ST.26 requires inclusion in a sequence listing of only the single sequence that has been enumerated by its residues. Therefore, at least one sequence containing any of 2, 3, or 100 copies of GGGX must be included in the sequence listing; however, the most encompassing sequence containing 100 copies of GGGX is preferred (SEQ ID NO: 42) (see Introduction to this document). In the latter case, a further annotation could indicate that up to 98 copies of GGGX could be deleted. Inclusion of two additional sequences containing 2 and 3 copies of GGGX, respectively (SEQ ID NO: 44-45), is strongly encouraged.

Further, the example does not disclose that the “z” variable is the same in each of the two occurrences in the expanded sequence. However, if “z” is disclosed as the same amino acid in all locations, then a feature Key VARIANT and a Qualifier NOTE should indicate that “X” in all positions can be any amino acid, as long as they are the same in all locations.

CAUTION: The preferred representation of the sequence indicated above is directed to the provision of a sequence listing on the filing date of a patent application. The same representation may not be applicable to a sequence listing provided subsequent to the filing date of a patent application, since consideration must be given to whether the information provided could be considered by an IPO to add subject matter to the original disclosure.

Relevant ST.26 paragraph(s): Paragraphs 3(c), 7(b), 26, and 27
Paragraph 28 – Amino acid sequences separated by internal terminator symbols

Example 28-1: Encoding nucleotide sequence and encoded amino acid sequence

A patent application describes the following sequences:

```
caatccaggg tggtgaat atg gcg ccc aat acg caa acc gcc tct ccc ggc
  Met Ala Pro Asn Thr Gln Thr Ala Ser Pro Arg
```

```
gcg ttg gcd gat tca tta atg cag ctg gca cga cag gtt tcc cga ctg
  Ala Leu Ala Asp Ser Leu Met Gln Leu Ala Arg Gln Val Ser Arg Leu
```

```
gaa agc ggg cag tga atg acc atg att acc gat tca ctc gcc gtc gtt
  Glu Ser Gly Gln   Met Thr Met Ile Thr Asp Ser Leu Ala Val Val
```

```
tta cca cgt cgt gac tgg gaa acc cct gcc gtt acc caa ctt aat cgc
  Leu Gln Arg Arg Asp Trp Glu Asn Pro Gly Val Thr Gln Leu Asn Arg
```

```
ctt gca gca cat tgg tgt caa aaa taa taataacgg atgtactatt
  Leu Ala Ala His Trp Cys Gln Lys
```

```
tatcctg atg ctg cgt cgt cag gtc aat gaa gtc gct taa gcaatcatg
  Met Leu Arg Arg Gln Val Asn Glu Val Ala
```

```
tcgatgcgg cgcgacgctt atccgaccaa catatcataa
```

Protein A

Protein B

Protein C

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

The application describes a nucleotide sequence containing termination codons, which encodes three distinct amino acids sequences.

The enumerated nucleotide sequence contains more than 10 specifically defined nucleotides and must be included in a sequence listing as a single sequence.

Regarding the encoded amino acid sequences, paragraph 28 requires that amino acid sequences separated by an internal terminator symbol such as a blank space, must be included as separate sequences. Since each of “Protein A”, “Protein B”, and “Protein C” contain four or more specifically defined amino acids, ST.26 paragraph 7(b) requires that each must be included in a sequence listing and must be assigned its own sequence identification number.
Question 3: How should the sequence(s) be represented in the sequence listing?

The nucleotide sequence must be included in a sequence listing as:

caattcagggtggtgaatatggcgcccaatacgcaaaccgcctctccccgcgcgttggccgattcattaatggaaagcgggcagtgaatgaccatgattacgga
 ttcac
tggccgtcgttttacaacgtcgtgactgggaaaaccctggcgttacccaacttaatcgccttgcagcacattggtgtcaaaaataataataaccggatgtact
 atttalcccgagtcgtgcgttcaggtgaatgaatcgccttaagcaatcaatgtcggatgcggcgcgacgcttatccgaccaacatatcataa. (SEQ ID NO: 46)

The nucleotide sequence should further be described using a “CDS” feature key for each of the three proteins and the element INSDFeature_location should identify the location of each coding sequence, including the stop codon. In addition, for each “CDS” feature key, the “translation” qualifier should be included with the amino acid sequence of the protein as the qualifier value. The application does not disclose the genetic code table that applies to the translation (see Annex 1, Section 9, Table 5). If the Standard Code table applies, then the qualifier “transl_table” is not necessary; however, if a different genetic code table applies, then the appropriate qualifier value from Table 5 must be indicated for the qualifier “transl_table”. Finally, the qualifier “protein_id” must be included with the qualifier value indicating the sequence identification number of each of the translated amino acid sequences.

The amino acid sequences must be included as separate sequences, each assigned its own sequence identification number:

MAPNTQTASPRALADSLMQLA RQVSRLESGQ (SEQ ID NO: 47)

MTMITDSLAVVLQRRDWENPGVTQLNRLAAHWCQK (SEQ ID NO: 48)

MLRQVNEVA (SEQ ID NO: 49)

NOTE: See “Example 90-1 Amino acid sequence encoded by a coding sequence with introns” for an illustration of a translated amino acid sequence represented as a single sequence.

Relevant ST.26 paragraphs: Paragraphs 7, 26, 28, 57, 87-90
Paragraph 29 – Representation of an “other” amino acid

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid

A patent application describes the following sequence:

\[ \text{Ala-Hse-X}_1\text{-X}_2\text{-X}_3\text{-X}_4\text{-Tyr-Leu-Gly-Ser} \]

Wherein, \( X_1 = \text{Ala or Gly} \),

\[ X_2 = \text{Ala or Gly} \]

\[ X_3 = \text{Ala or Gly} \]

\[ X_4 = \text{Ala or Gly} \]

and

\[ \text{Hse = Homoserine} \]

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

The enumerated peptide contains five specifically defined amino acids. The symbol “X” is used conventionally to represent two amino acids in the alternative (see Introduction to this document).

Because there are five specifically defined amino acids, i.e., Ala, Tyr, Leu, Gly and Ser, ST.26 paragraph 7(b) requires that the sequence must be included in a sequence listing.

Question 3: How should the sequence(s) be represented in the sequence listing?

Paragraph 29 requires any “other” amino acid must be represented by the symbol “X”. In the example, the sequence contains the amino acid Hse in position 2 which is not found in Annex I, Section 3, Table 3. Accordingly, Hse is an “other” amino acid and must be represented by the symbol “X”.

\( X_1\text{-X}_4 \) are variant positions, each of which can be A or G. The most restrictive ambiguity symbol for alternatives A or G is “X”. Therefore, the sequence may be represented as:

\[ \text{AXXXXXYLGSS} \text{ (SEQ ID NO: 50)} \]

Inclusion of any specific sequences essential to the disclosure or claims of the invention is strongly encouraged, as discussed in the introduction to this document.

Since amino acid Hse is not found in Annex I, Section 4, Table 4, a feature key “SITE” and a qualifier “NOTE” must be provided with the complete, unabbreviated name of Homoserine.

According to paragraph 27, because \( X_1\text{-X}_4 \) represent an alternative of only 2 amino acids, then further description is required. Paragraph 94 indicates that the feature key “VARIANT” should be used with the qualifier “NOTE” and qualifier value “A or G”. According to ST.26 paragraph 34, since these positions are adjacent and have the same description, they may be jointly described using the syntax “3..6” as the location descriptor in the element INSDFeature_location.

Relevant ST.26 paragraphs: Paragraphs 3(a), 7(b), 25-27, 29, 34, 66, 70, 71, and 94-95
Paragraph 30 – Annotation of a modified amino acid

Example 30-1 – Feature key “CARBOHYD”

A patent application describes a polypeptide with a specifically modified amino acid, containing a glycosylated side chain, characterized in that Cys corresponding to positions 4 and 15 of the polypeptide forms a disulfide bond, according to the following sequence:

Leu-Glu-Tyr-Cys-Leu-Lys-Arg-Trp-Asn(asialyloligosaccharide)-Glu-Thr-Ile-Ser-His-Cys-Ala-Trp

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

The enumerated peptide provides 17 specifically defined amino acids. There are 16 natural amino acids, wherein the ninth (asparagine) is glycosylated. Therefore, the sequence must be included in a sequence listing as required by ST.26 paragraph (7)(b).

Question 3: How should the sequence(s) be represented in the sequence listing?

According to ST.26 paragraph 29, a modified amino acid should be represented in the sequence as the corresponding unmodified amino acid whenever possible.

Therefore the sequence must be included in a sequence listing as:

LEYCLKRWNETISHCAW (SEQ ID NO: 51)

A further description of the modified amino acid is required. The feature key “CARBOHYD” together with the (mandatory) qualifier “NOTE” should be used to indicate the occurrence of the attachment of a sugar chain (asialyloligosaccharide) to asparagine in position 9. The qualifier “NOTE” describes the type of linkage, e.g. N-linked. The location descriptor in the feature location element is the residue position number of the modified asparagine.

In addition, there is a disulfide bond between the two Cys residues. Therefore the feature key “DISULFID” is used to describe an intrachain crosslink. The location descriptors in the feature location element are the residue position numbers of the linked Cys residues in conjunction with the “join” location operator, “join(4,15)”. The qualifier NOTE is not mandatory.

Relevant ST.26 paragraph(s): Paragraphs 3(a), 7(b), 26, 29, 30, and Annex I, section 7, feature key 7.4
Paragraph 36 – Sequences containing regions of an exact number of contiguous "n" or "X" residues

Example 36-1: Sequence with a region of a known number of "X" residues represented as a single sequence

LL-100-KYMR

Where the "-100-" between amino acids Leucine and Lysine reflects a 100 amino acid region in the sequence.

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

ST.26 paragraph 36 requires inclusion of a sequence that contains at least four specifically defined amino acids separated by one or more regions of a defined number of "X" residues.

The disclosed sequence uses a nonconventional symbol, i.e. "-100-". The definition of "-100-" must be determined from the explanation of the sequence in the disclosure, which defines this symbol as 100 amino acids between leucine and lysine (see Introduction to this document). Therefore, "-100-" is a defined region of "X" residues. Since six of the 106 amino acids in the sequence are specifically defined, ST.26 paragraph 7(b) requires that the sequence must be included in a sequence listing.

Question 3: How should the sequence(s) be represented in the sequence listing?

The nonconventional symbol "-100-" is represented as 100 "X" residues (since any symbol used to represent an amino acid is equivalent to only one residue). Therefore, a single sequence of 106 amino acids in length containing 100 "X" residues between LL and KYMR must be included in a sequence listing (SEQ ID NO: 52).

Relevant ST.26 paragraph(s): Paragraphs 7(b), 26, 27, and 36
Example 36-2: Sequence with multiple regions of a known number or range of "X" residues represented as a single sequence

Lys-z\textsubscript{2}-Lys-z\textsubscript{3}-Lys-z\textsubscript{2}-Lys-z\textsubscript{2}-Lys

Where z is any amino acid, m=20, n=19-20, z\textsubscript{2} means that the pairs of Lysines are separated by any two amino acids, and z\textsubscript{3} means the pairs of Lysines are separated by any three amino acids.

**Question 1:** Does ST.26 require inclusion of the sequence(s)?

**YES**

The disclosed sequence uses a nonconventional symbol, i.e. "z." Therefore, the disclosure must be consulted to determine the definition; "z" is defined as any amino acid (see Introduction to this document). The conventional symbol used to represent any amino acid is "X." Considering the presence of "X" variables, the peptide contains six lysine residues that are enumerated and specifically defined, which is required to be included in a sequence listing.

**Question 3:** How should the sequence(s) be represented in the sequence listing?

The sequence uses a nonconventional symbol "z," the definition of which must be determined from the disclosure. Since "z" is defined as any amino acid, the conventional symbol is "X."

The preferred and most encompassing means of representation is (see Introduction to this document):

$$\text{KXXKXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXKXXK}$$ (SEQ ID NO: 53)

Wherein z\textsubscript{n} is equal to 20 "X"s, with a further description that the "X" variable corresponding to position 30 can be deleted.

Alternatively, or in addition to the above, the sequence may be represented as:

$$\text{KXXKXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXKXXK}$$ (SEQ ID NO: 54)

Wherein z\textsubscript{n} is equal to 19 "X"s; with a further description that an "X" variable between position numbers 29 and 30 can be inserted.

**Relevant ST.26 paragraph(s):** Paragraphs 26, 27, and 36
Example 36-3: Sequence with multiple regions of a known number or range of “X” residues represented as a single sequence

K_{z_2}K_{z_0}K_{z_5}K_{z_0}K_{z_2}K

Where z is any amino acid, where m=15-25, preferably 20-22, n=15-25, preferably 19-20, z_2 means that the pairs of Lysines are separated by any two amino acids, and z_5 means the pairs of Lysines are separated by any three amino acids.

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

The sequence in the example uses a nonconventional symbol, i.e. “z.” Therefore, the surrounding disclosure is consulted to determine the definition of “z” (see Introduction to this document). The disclosure defines this symbol as any amino acid. The conventional symbol used to represent this amino acid is “X.” After considering the presence of “X” variables, the peptide contains 6 lysine residues that are enumerated and specifically defined, which is required in a sequence listing.

Question 3: How should the sequence(s) be represented in the sequence listing?

The sequence uses a nonconventional symbol “z,” the definition of which must be determined from the disclosure. Since “z” is defined as any amino acid, the conventional symbol is “X.” The preferred and most encompassing means of representation is:

KXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXK (SEQ ID NO: 55)

(where m=25 and n=25), with a further description that up to 10 “X” residues in each of the “z_0” or “z_5” regions may be deleted.

Inclusion of any specific sequences essential to the disclosure or claims of the invention is strongly encouraged, as discussed in the introduction to this document.

Alternatively, the sequence may be represented as:

KXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXK (SEQ ID NO: 56)

(where m=15 and n=15), with a further description that up to 10 “X” residues in each of the “z_0” or “z_5” regions may be inserted.

As further alternatives, any or all possible variations may be included.

CAUTION: The preferred representation of the sequence indicated above is directed to the provision of a sequence listing on the filing date of a patent application. The same representation may not be applicable to a sequence listing provided subsequent to the filing date of a patent application, since consideration must be given to whether the information provided could be considered by an IPO to add subject matter to the original disclosure.

Relevant ST.26 paragraph(s): Paragraphs 27 and 36
Paragraph 37 – Sequences containing regions of an unknown number of “n” or “X” residues

Example 37-1: Sequence with regions of an unknown number of “X” residues must not be represented as a single sequence

Gly-Gly----Gly-Gly-Xaa-Xaa

where the symbol ---- is an undefined gap within the sequence, where Xaa is any amino acid, and the Glycine and Xaa residues are connected to one another through peptide bonds.

Question 1: Does ST.26 require inclusion of the sequence(s)?

NO

ST.26 paragraph 37 prohibits the inclusion of any sequence that contains an undefined gap; therefore, inclusion of the entire sequence is not required.

ST.26 paragraph 37 does require inclusion of any portion of a sequence adjacent to an undefined gap that contains four or more specifically defined amino acids. In the example above, inclusion of either portion adjacent to the undefined gap is not required, since each portion contains only two specifically defined amino acids.

Question 2: Does ST.26 permit inclusion of the sequence(s)?

NO – not the entire sequence

NO – not any portion of the sequence

ST.26 paragraph 37 does not permit inclusion of the entire sequence.

ST.26 paragraph 37 does not permit inclusion of either portion adjacent to the undefined gap, since each portion contains only two specifically defined amino acids.

Relevant ST.26 paragraphs: Paragraphs 7(b), 8, 26, and 37
Example 37-2: Sequence with regions of an unknown number of “X” residues must not be represented as a single sequence

Gly-Gly----Gly-Gly-Ala-Gly-Xaa-Xaa

wherein the symbol ---- is an undefined gap within the sequence, where Xaa is any amino acid, and the Glycine and Xaa residues are connected to one another through peptide bonds.

Question 1: Does ST.26 require inclusion of the sequence(s)?

NO – not the entire sequence

YES – a portion of the sequence

ST.26 paragraph 37 prohibits the inclusion of any sequence that contains an undefined gap, but requires inclusion of any portion of a sequence adjacent to an undefined gap that contains four or more specifically defined amino acids.

In the example above, ST.26 does not require (and prohibits) inclusion of both the entire sequence, which contains an undefined gap, and the Gly-Gly portion adjacent to the undefined gap, which contains only two specifically defined amino acids. However, ST.26 requires inclusion of the Gly-Gly-Ala-Gly-Xaa-Xaa portion adjacent to the undefined gap, since it contains at least four specifically defined amino acids.

Question 2: Does ST.26 permit inclusion of the sequence(s)?

NO – not the entire sequence and not the Gly-Gly portion

Question 3: How should the sequence(s) be represented in the sequence listing?

The portion of the sequence adjacent to the undefined gap that contains four specifically defined amino acids must be represented as:

GGAGXX (SEQ ID NO: 57)

Preferably, the sequence should be annotated to indicate that the represented sequence is part of a larger sequence that contains an undefined gap by using the feature key “SITE”, the feature location “1” and the qualifier “NOTE” with the value, e.g., “This residue is linked N-terminally to a peptide having an N-terminal Gly-Gly and a gap of undefined length.”

Relevant ST.26 paragraph(s): Paragraphs 7(b), 8, 26, and 37
Paragraph 87 – “CDS” Feature key

Example 87-1: Encoding nucleotide sequence and encoded amino acid sequence

A patent application describes the following nucleotide sequence and its translation:

\[
\begin{align*}
\text{atg acc gga aat aaa cct gaa acc gat gtt tac gaa att tta tga} \\
\text{Met Thr Gly Asn Lys Pro Glu Thr Asp Val Tyr Glu Ile Leu STOP}
\end{align*}
\]

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES – the nucleotide sequence. The enumerated nucleotide sequence has more than ten specifically defined nucleotides.

YES – the peptide sequence. The enumerated peptide sequence has more than four specifically defined amino acids.

Question 3: How should the sequence(s) be represented in the sequence listing?

The nucleotide sequence must be presented as:

\[
\text{atgaccgaaatatggtttacgaaattttatga (SEQ ID NO: 58)}
\]

The nucleotide sequence should further be described using the “CDS” feature key and the element INSDFeature_location should identify the entire sequence, including the stop codon (i.e., position 1 through 45).

In addition, the “translation” qualifier should be included with the qualifier value “MTGNKPETDVYEIL”. The application does not disclose the genetic code table that applies to the translation (see Annex 1, Section 9, Table 5). If the Standard Code table applies, then the qualifier “transl_table” is not necessary; however, if a different genetic code table applies, then the appropriate qualifier value from Table 5 must be indicated for the qualifier “transl_table”. Finally, the qualifier “protein_id” must be included with the qualifier value indicating the sequence identification number of the translated peptide.

The peptide sequence must be separately presented with its own sequence identification number using single letter codes as follows:

\[
\text{MTGNKPETDVYEIL (SEQ ID NO: 59)}
\]

The STOP following the enumerated peptide sequence must not be included in the peptide sequence in the sequence listing.

CAUTION: The preferred representation of the sequence indicated above is directed to the provision of a sequence listing on the filing date of a patent application. The same representation may not be applicable to a sequence listing provided subsequent to the filing date of a patent application, since consideration must be given to whether the information provided could be considered by an IPO to add subject matter to the original disclosure.

Relevant ST.26 paragraphs: Paragraphs 7(a), 7(b), 26, 28, 87, 88, and 90
Paragraph 90 – Amino acid sequence encoded by a coding sequence

Example 90-1: Amino acid sequence encoded by a coding sequence with introns

A patent application contains the following figure disclosing a coding sequence and its translation:

```
Figure 1 – nucleotides shown in bold-face are intron regions.
```

```
Paragraph 90 – Amino acid sequence encoded by a coding sequence

Example 90-1: Amino acid sequence encoded by a coding sequence with introns

A patent application contains the following figure disclosing a coding sequence and its translation:

```
Paragraph 90 – Amino acid sequence encoded by a coding sequence

Example 90-1: Amino acid sequence encoded by a coding sequence with introns

A patent application contains the following figure disclosing a coding sequence and its translation:

```
```

```
The nucleotide sequence contains coding sequence (exons) separated by noncoding sequence (introns). According to the figure caption, the bolded regions of nucleotides are intron sequences that will be spliced out of an RNA transcript before translation into a protein. Accordingly, the three amino acid sequences are actually a single, contiguous, enumerated sequence, which contains more than four specifically defined amino acids and must be included in a sequence listing as a single sequence.
```

Question 3: How should the sequence(s) be represented in the sequence listing?

The nucleotide sequence must be included in a sequence listing as:

```
atgaagactttcgcagccttgctttccgctgtctcgtcactctcgcgctctcggtgcgcgcccaggcggctgtctggagtcaatgtaagtgccgctgcttttcttgcttgacgtgcgtttgtacctccttgagctcagtga
cggactttcaatccgtcgtcattgctcctcatgtattgacgattggccttcatagtcatactctcaatgcgttccgggctccgcaacgtccagcgctccggcggccccctcagcgacaacttcaggccccgcacctacggacggaacgtgctcggccagcggggcatggccgccattgacctga (SEQ ID NO: 74)
```
The nucleotide sequence should further be described using a “CDS” feature key and the element “NSDFeature, location” should identify the location of the coding sequence, including the stop codon indicated by “Ter”. In addition, the “translation” qualifier should be included, with the amino acid sequence of the protein as the qualifier value. (Note that the terminator symbol “Ter” in the last position of the sequence must not be included in the amino acid sequence.) The application does not disclose the genetic code table that applies to the translation (see Annex 1, Section 9, Table 5). If the “Standard Code” table applies, then the qualifier “transl_table” is not necessary; however, if a different genetic code table applies, then the appropriate qualifier value from Table 5 must be indicated for the qualifier “transl_table”. Finally, the qualifier “protein_id” must be included with the qualifier value indicating the sequence identification number of the translated amino acid sequence.

The amino acid sequence must be included as a single sequence:

MKTFAALLSVAATLALSVRAAQAVWSQCQGTPGWGTTTCVAGSVCTSLSSYYSQCVPGSATSSAPAAPSATTSGAPTDGTCGSGAWPLT (SEQ ID NO: 76)

Relevant ST.26 paragraphs: Paragraphs 7, 26, 28, 57, 87-90
**Paragraph 91 – Primary sequence and a variant, each enumerated by its residues**

**Example 91-1: Representation of enumerated variants**

The description includes the following sequence alignment:

- **D. melanogaster**: `ACATTGAATCTCATACCACTTT`
- **D. virilis**: `...---G...G...G.....`
- **D. simulans**: `GT..G.CG..GT..GT.G...`

**Question 1: Does ST.26 require inclusion of the sequence(s)?**

**YES**

It is common in the art to include “dots” in a sequence alignment to indicate “this position is the same as the position above it.” Therefore, the “dots” in species 2 and 3 are considered enumerated and specifically defined nucleotides, as they are simply a short-hand way of indicating that a given position is the same nucleotide as in species 1. In addition, sequence alignments frequently display the symbol “-” to indicate the absence of a residue in order to maximize the alignment.

Accordingly, the nucleotide sequences of species 1 and 3 contain twenty-two enumerated and specifically defined nucleotides, whereas the nucleotide sequences of species 2 contains nineteen. Thus, each sequence is required by ST.26 paragraph 7(a) to be included in a sequence listing with separate sequence identification numbers.

**Question 3: How should the sequence(s) be represented in the sequence listing?**

- **Drosophila melanogaster** sequence must be included in a sequence listing as: `acattgaatctcatacctt` (SEQ ID NO: 60)
- **Drosophila virilis** sequence must be included in a sequence listing as: `acattgaatctcatacctt` (SEQ ID NO: 61)
- **Drosophila simulans** sequence must be included in a sequence listing as: `gtatggcgtcgtatsgtagttt` (SEQ ID NO: 62)

**Relevant ST.26 paragraphs:** Paragraphs 7(a), 13, and 91
Example 91-2: Representation of enumerated variants

The description includes the following table of a peptide and functional variants thereof. A blank space in the table below indicates that an amino acid in the variant is the same as the corresponding amino acid in the "Sequence" and a "-" indicates deletion of the corresponding amino acid in the "Sequence".

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequence</strong></td>
<td>A</td>
<td>V</td>
<td>L</td>
<td>L</td>
<td>Y</td>
<td>L</td>
<td>R</td>
<td>G</td>
<td>E</td>
</tr>
<tr>
<td><strong>Variant 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td></td>
</tr>
<tr>
<td><strong>Variant 2</strong></td>
<td></td>
<td>P</td>
<td></td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Variant 3</strong></td>
<td>A</td>
<td>I</td>
<td>G</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Variant 4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>

**Question 1: Does ST.26 require inclusion of the sequence(s)?**

**YES**

As indicated, a blank space in this table indicates that an amino acid in the variant is the same as the corresponding amino acid in the "Sequence". Therefore, the amino acids of the variant sequences are enumerated and specifically defined.

Since the four variant sequences each contain more than four enumerated and specifically defined amino acids, each sequence is required by ST.26 paragraph 7(a) to be included in a sequence listing with separate sequence identification numbers.

**Question 3: How should the sequence(s) be represented in the sequence listing?**

AVLTYLRGE (SEQ ID NO: 76)
AVLTYLRGA (SEQ ID NO: 77)
AVPTYPRGE (SEQ ID NO: 78)
AVAIGYRGE (SEQ ID NO: 79)
AVLTYLGE (SEQ ID NO: 80)

**Relevant ST.26 paragraphs:** Paragraphs 7(b), 26, and 91
Example 91-3: Representation of a consensus sequence

A patent application includes Figure 1 with the following multiple sequence alignment.

Consensus:  LEQNEQFINAAKIIRHPYnrrk'lNND1MLIK

Homo sapiens:  LEQNEQFINAAKIIRHPYDRKT1NND1MLIK

Pongo abelii:  LEQNEQFINAAKIIRHPYDRKTVND1MLIK

Papio Anubis:  LEQNEQFINAAKIIRHPYDRKTVND1MLIK

Rhinopithecus roxellana:  LEQNEQFINAAKIIRHPYNRITLND1MLIK

Pan paniscus:  LEQNEQFINAAKIIRHPYNRITLND1MLIK

Rhinopithecus bieti:  LEQNEQFINAAKIIRHPYN1NLND1MLIK

Rhinopithecus roxellana:  LEQNEQFINATQ1IRHPYN1NLND1MLIK

The consensus sequence includes upper case letters to represent conserved amino acid residues, while the lower case letters "n", "a", "k", "r", "i" and "m" represent the predominant amino acid residues among the aligned sequences.

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

The lower case letters in the consensus sequence each represent a single amino acid residue. Consequently, the consensus sequence, as well as each of the remaining seven sequences in Figure 1, includes at least four specifically defined amino acids. ST.26 paragraph 7(b) requires inclusion of all eight sequences in the sequence listing.

Question 3: How should the sequence(s) be represented in the sequence listing?

The lower case letters in the consensus sequence are being used as ambiguity symbols to represent the predominant amino acid among the possible variants for a specific position. Therefore, the lower case letters "n", "a", "k", "r", "i" and "m" are conventional symbols used in a nonconventional manner and the consensus sequence must be represented using an ambiguity symbol in place of each of the lower case letters.

The most restrictive ambiguity symbol should be used. For most positions in the consensus sequence, "X" is the most restrictive ambiguity symbol; however, the most restrictive ambiguity symbol for "D" or "N" in positions 20 and 25 is "B". The consensus sequence should be included in the sequence listing as:

LEQNEQFINAXXIRHPYBXXTXBNDIXLIK (SEQ ID NO: 81)

According to paragraph 27, the symbol "X" will be construed as any one of "A", "R", "N", "D", "C", "Q", "E", "G", "H", "I", "L", "K", "M", "F", "P", "O", "S", "U", "T", "W", "Y", or "V", except where it is used with a further description in the feature table. Therefore, each "X" in the consensus sequence must be further described in a feature table using the feature key "VARIANT" and the qualifier "NOTE" to indicate the possible variants for each position.

The remaining seven sequences must be included in the sequence listing as:

LEQNEQFINAAKIIRHPYDRKT1NND1MLIK (SEQ ID NO: 82)

LEQNEQFINAAKIIRHPYDRKTVND1MLIK (SEQ ID NO: 83)

LEQNEQFINAAKIIRHPYDRKT1NND1MLIK (SEQ ID NO: 84)

LEQNEQFINAAKIIRHPYDRKT1NND1MLIK (SEQ ID NO: 85)

LEQNEQFINAAKIIRHPKYN1NITLND1MLIK (SEQ ID NO: 86)

LEQNEQFINATQ1IRHPYN1NLND1MLIK (SEQ ID NO: 87)

LEQNEQFINATQ1IRHPYN1NLND1MLIK (SEQ ID NO: 88)

CAUTION: The preferred representation of the sequence indicated above is directed to the provision of a sequence listing on the filing date of a patent application. The same representation may not be applicable to a sequence listing provided subsequent to the filing date of a patent application, since consideration must be given to whether the information provided could be considered by an IPO to add subject matter to the original disclosure.

Relevant ST.26 paragraphs: Paragraphs 7(b), 26, 27, 91, and 95
Paragraph 92 – Variant sequence disclosed as a single sequence with enumerated alternative residues

Example 92-1: Representation of single sequence with enumerated alternative amino acids

A patent application claims a peptide of the sequence:

(i) Gly–Gly–Gly–[Leu or Ile]–Ala–Thr–[Ser or Thr]

Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

The sequence provides four specifically defined amino acids and ST.26 paragraph 7(b) requires inclusion of the sequence in a sequence listing.

Question 3: How should the sequence(s) be represented in the sequence listing?

Table 3 of Annex I, Section 3 defines the ambiguity symbol “J” as isoleucine or leucine. Therefore, the preferred representation of the sequence is:

GGGJATX (SEQ ID NO: 63)

which requires a further description in a feature table using the feature key “VARIANT” and the qualifier “NOTE” to indicate that the “X” is Serine or Threonine.

Alternatively, the sequence may be represented, for example, as:

GGGLATS (SEQ ID NO: 64)

which requires a further description in a feature table using the feature key “VARIANT” and the qualifier “NOTE” to indicate that L can be replaced by I, and S can be replaced by T.

CAUTION: The preferred representation of the sequence indicated above is directed to the provision of a sequence listing on the filing date of a patent application. The same representation may not be applicable to a sequence listing provided subsequent to the filing date of a patent application, since consideration must be given to whether the information provided could be considered by an IPO to add subject matter to the original disclosure.

Relevant ST.26 paragraph(s): Paragraphs 7(b), 8, 26, 27, 92, and 95
**Paragraph 93(a)** – A variant sequence disclosed only by reference to a primary sequence with multiple independent variations.

**Example 93(a)-1: Representation of a variant sequence by annotation of the primary sequence**

An application contains the following disclosure:

*Peptide fragment 1 is Gly-Leu-Pro-Xaa-Arg-Ile-Cys wherein Xaa can be any amino acid...*

In another embodiment, peptide fragment 1 is Gly-Leu-Pro-Xaa-Arg-Ile-Cys wherein Xaa can be Val, Thr, or Asp...

In another embodiment, peptide fragment 1 is Gly-Leu-Pro-Xaa-Arg-Ile-Cys wherein Xaa can be Val.

**Question 1:** Does ST.26 require inclusion of the sequence(s)?

**YES**

"Peptide fragment 1" in each of the three disclosed embodiments provides at least six specifically defined amino acids; therefore, the sequence must be included in a sequence listing as required by ST.26 paragraph 7(b).

**Question 3:** How should the sequence(s) be represented in the sequence listing?

In this example, the enumerated sequence of "Peptide fragment 1" is disclosed three times, as three different embodiments, each with an alternative description of Xaa. In this example, "X" is the most restrictive ambiguity symbol for the Xaa position.

ST.26 requires inclusion of the disclosed enumerated sequence only once. In the most encompassing of the three embodiments, Xaa is any amino acid (see Introduction to this document). Therefore, the sequence that must be included in the sequence listing is:

GLPXRIC (SEQ ID NO: 65)

Inclusion of any additional sequences essential to the disclosure or claims of the invention is strongly encouraged, as discussed in the introduction to this document.

For the above example, it is strongly encouraged that the following additional three sequences are included in the sequence listing, each with their own sequence identification number:

GLPVRIC (SEQ ID NO: 66)

GLPTRIC (SEQ ID NO: 67)

GLPDRIC (SEQ ID NO: 68)

**CAUTION:** The preferred representation of the sequence indicated above is directed to the provision of a sequence listing on the filing date of a patent application. The same representation may not be applicable to a sequence listing provided subsequent to the filing date of a patent application, since consideration must be given to whether the information provided could be considered by an IPO to add subject matter to the original disclosure.

**Relevant ST.26 paragraph(s):** Paragraphs 7(b), 26, 27, and 93(a)
**Paragraph 93(b)** – A variant sequence disclosed only by reference to a primary sequence with multiple interdependent variations

**Example 93(b)-1:** Representation of individual variant sequences with multiple interdependent variations

A patent application describes the following consensus sequence:

```
cgaatg
```

wherein \( n_1, n_2, \) and \( n_3 \) can be a, t, g, or c.

Several variant sequences are disclosed as follows:

- If \( n_1 \) is a, then \( n_2 \) and \( n_3 \) are t, g, or c;
- If \( n_1 \) is t, then \( n_2 \) and \( n_3 \) are a, g, or c;
- If \( n_1 \) is g, then \( n_2 \) and \( n_3 \) are t, a, or c;
- If \( n_1 \) is c, then \( n_2 \) and \( n_3 \) are t, g, or a.

**Question 1:** Does ST.26 require inclusion of the sequence(s)?

**YES**

The sequence has more than ten enumerated and “specifically defined” nucleotides and is required by ST.26 paragraph 7(a) to be included in a sequence listing.

**Question 3:** How should the sequence(s) be represented in the sequence listing?

The enumerated sequence contains more than ten specifically defined nucleotides and three “n” residues. ST.26 requires inclusion of the disclosed enumerated sequence and where an ambiguity symbol is appropriate, the most restrictive symbol should be used. In this example, \( n_1, n_2, \) and \( n_3 \) can be a, t, g, or c, so “n” is the most restrictive ambiguity symbol. Therefore, the sequence that must be included in the sequence listing is:

```
cgaatgncccactacgaatgnacgaatgncccaca
```

The enumerated sequence contains variations at three distinct locations and the occurrence of the variations is interdependent. Inclusion of additional sequences which represent additional embodiments that are a key part of the invention is strongly encouraged, as discussed in the introduction to this document. Therefore, according to ST.26 paragraph 93(b), the additional embodiments should be included in a sequence listing as four separate sequences, each with its own sequence identification number:

```
cgaatgacccactacgaatgbcacgaatgbcccaca (SEQ ID NO: 69)
cgaatgtcccactacgaatgvcacgaatgvcccaca (SEQ ID NO: 70)
cgaatgccccactacgaatgbacgaatgbcccaca (SEQ ID NO: 71)
cgaatgccccactacgaatqdcacgaatqdcgccaca (SEQ ID NO: 72)
```

(Note that \( b = t, g, \) or c; \( v = a, g, \) or c; \( h = t, a, \) or c; and \( d = t, g, \) or a; see Annex I, Section 1, Table 1)

According to ST.26 paragraph 15, the most restrictive symbol must be used to represent variable positions. Consequently, \( n_2 \) and \( n_3 \) must not be represented by “n” in the sequence.

**CAUTION:** The preferred representation of the sequence indicated above is directed to the provision of a sequence listing on the filing date of a patent application. The same representation may not be applicable to a sequence listing provided subsequent to the filing date of a patent application, since consideration must be given to whether the information provided could be considered by an IPO to add subject matter to the original disclosure.

**Relevant ST.26 paragraphs:** Paragraphs 7(a), 15, and 93(b)

[Appendix of Annex VI follows]
APPENDIX

GUIDANCE DOCUMENT SEQUENCES IN XML:

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ST26SequenceListing PUBLIC "-//WIPO/DTD Sequence Listing 1.0//EN" "resources/ST26SequenceListing_V1_0.dtd">
<ST26SequenceListing version="1.0" productiondate="2017-01-02">
  <ApplicantFileReference>ABCD1234</ApplicantFileReference>
  <ApplicantName languageCode="de">Johannes Jäger</ApplicantName>
  <ApplicantNameLatin>Pharmakologische Wirkstoffe für das Nervensystem</ApplicantName>
  <InventionTitle>Pharmakologische Wirkstoffe für das Nervensystem</InventionTitle>
  <SequenceTotalQuantity>88</SequenceTotalQuantity>
  <SequenceData sequenceIDNumber="1">
    <INSDSeq length="7"/>
    <INSDSeq moltype="AA"/>
    <INSDSeq division/>
    <INSDSeq feature/>
    <INSDFeature key="SOURCE"/>
    <INSDFeature location>1..7</INSDFeature location>
    <INSDFeature key="ORGANISM"/>
    <INSDQualifier name="Homo sapiens"/>
    <INSDQualifier value="protein"/>
    <INSDQualifier value="circular peptide"/>
    <INSDQualifier name="MOL_TYPE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>
    <INSDQualifier name="NOTE"/>...<INSDQualifier name="NOTE"/>
  </SequenceData>
</ST26SequenceListing>
```

Date: September 2017
<INSDSeq>
  <INSDseq_sequence>VAFK/INSDES sequence</INSDseq_sequence>
  <INSDseq_length>5</INSDseq_length>
  <INSDseq_moltype>AA</INSDseq_moltype>
  <INSDseq_division>PAT</INSDseq_division>
  <INSDseq_feature-table>
    <INSDfeature>
      <INSDfeature_key>SOURCE</INSDfeature_key>
      <INSDfeature_location>1..5</INSDfeature_location>
      <INSDQualifier>
        <INSDQualifier_name>ORGANISM</INSDQualifier_name>
        <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
      </INSDQualifier>
      <INSDQualifier>
        <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
        <INSDQualifier_value>protein</INSDQualifier_value>
      </INSDQualifier>
    </INSDfeature>
    <INSDfeature>
      <INSDfeature_key>SEQ_ID</INSDfeature_key>
      <INSDfeature_location>1..5</INSDfeature_location>
      <INSDQualifier>
        <INSDQualifier_name>UNG</INSDQualifier_name>
        <INSDQualifier_value>the N-terminus is acetylated</INSDQualifier_value>
      </INSDQualifier>
    </INSDfeature>
    <INSDfeature>
      <INSDfeature_key>FEATURE</INSDfeature_key>
      <INSDfeature_location>1..5</INSDfeature_location>
      <INSDQualifier>
        <INSDQualifier_name>NOTE</INSDQualifier_name>
        <INSDQualifier_value>C-terminus linked via a glutaraldehyde bridge to dipeptide GKK</INSDQualifier_value>
      </INSDQualifier>
    </INSDfeature>
  </INSDseq_feature-table>
</INSDSeq>
cytosine attached to a C3 spacer, which is joined to another nucleic acid. The entire sequence of amino acids 1-3 can be repeated one or more times.

Ref.: SequenceData
<INSDSeq_length>12</INSDSeq_length>
<INSDSeq_moltype>DNA</INSDSeq_moltype>
<INSDSeq_division>PAT</INSDSeq_division>
<INSDSeq_feature-tables>
<INSDFeature>
<INSDFeature_key>source</INSDFeature_key>
<INSDFeature_location>1..12</INSDFeature_location>
<INSDFeature_qual>
<INSDQualifier name=organism>
<INSDQualifier_value>other DNA</INSDQualifier_value>
</INSDQualifier>
</INSDFeature>
</INSDSeq_feature-tables>
</SequenceData>
<SequenceData_sequenceIDNumber="6" page: 3.26.vi.66>
<INSDSeq>
<INSDSeq_length>12</INSDSeq_length>
<INSDSeq_moltype>DNA</INSDSeq_moltype>
<INSDSeq_division>PAT</INSDSeq_division>
<INSDSeq_feature-tables>
<INSDFeature>
<INSDFeature_key>source</INSDFeature_key>
<INSDFeature_location>1..12</INSDFeature_location>
<INSDFeature_qual>
<INSDQualifier name=organism>
<INSDQualifier_value>other DNA</INSDQualifier_value>
</INSDQualifier>
</INSDFeature>
</INSDSeq_feature-tables>
</SequenceData>

<INSDQualifier name="organism" value="Homo sapiens"/>
<INSDQualifier name="mol_type" value="DNA"/>
<INSDQualifier name="other DNA" value="C3 spacer"/>
<INSDQualifier name="note" value="c"/>
<INSDQualifier name="note" value="other DNA"/>
<INSDFeature>
  <INSDFeature_key>misc_feature</INSDFeature_key>
  <INSDFeature_location>1..30</INSDFeature_location>
</INSDFeature>
<INSDQualifier name="note" value="end of another nucleic acid"/>
<INSDFeature>
  <INSDFeature_key>misc_feature</INSDFeature_key>
  <INSDFeature_location>1..18</INSDFeature_location>
</INSDFeature>
<INSDQualifier name="organism" value="Homo sapiens"/>
<INSDQualifier name="mol_type" value="DNA"/>
<INSDQualifier name="note" value="genic DNA"/>
<INSDQualifier name="note" value="basic site"/>
<INSDFeature>
  <INSDFeature_key>misc_feature</INSDFeature_key>
  <INSDFeature_location>11..18</INSDFeature_location>
</INSDFeature>
<INSDQualifier name="mod_base" value="cytosine"/>
<INSDFeature>
  <INSDFeature_key>misc_feature</INSDFeature_key>
  <INSDFeature_location>1..30</INSDFeature_location>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>misc_feature</INSDFeature_key>
  <INSDFeature_location>11..30</INSDFeature_location>
</INSDFeature>
<INSDQualifier name="note" value="industrial property"/>
<INSDQualifier name="note" value="other DNA"/>
<INSDQualifier name="note" value="organic"/>
<INSDQualifier name="note" value="synthetic construct"/>
<INSDQualifier name="note" value="ST.26"/>
<INSDQualifier name="note" value="modified_base"/>
<INSDQualifier name="note" value="HANDBOOK ON INFORMATION AND DOCUMENTATION"/>
<INSDQualifier name="note" value="INDUSTRIAL PROPERTY"/>
<INSDQualifier name="note" value="organism"/>

Date: September 2017

en / 03-26-01
<INSDFeature>
  <INSDFeature_key>modified base</INSDFeature_key>
  <INSDFeature_location>1..10</INSDFeature_location>
  <INSDFeature_qua>
    <INSDFeature_key>mod_base</INSDFeature_key>
    <INSDFeature_value>G</INSDFeature_value>
    <INSDFeature_value>C</INSDFeature_value>
    <INSDFeature_value>T</INSDFeature_value>
    <INSDFeature_value>A</INSDFeature_value>
    <INSDFeature_value>U</INSDFeature_value>
  </INSDFeature_qua>
</INSDFeature>

<INSDFeature>
  <INSDFeature_key>source</INSDFeature_key>
  <INSDFeature_location>1..10</INSDFeature_location>
  <INSDFeature_qua>
    <INSDFeature_key>organism</INSDFeature_key>
    <INSDFeature_value>other DNA</INSDFeature_value>
    <INSDFeature_value>RNA</INSDFeature_value>
    <INSDFeature_value>other RNA</INSDFeature_value>
    <INSDFeature_value>DNA</INSDFeature_value>
    <INSDFeature_value>RNA</INSDFeature_value>
  </INSDFeature_qua>
</INSDFeature>

<INSDFeature>
  <INSDFeature_key>mol_type</INSDFeature_key>
  <INSDFeature_value>other DNA</INSDFeature_value>
  <INSDFeature_value>other RNA</INSDFeature_value>
  <INSDFeature_value>RNA</INSDFeature_value>
  <INSDFeature_value>DNA</INSDFeature_value>
  <INSDFeature_value>other RNA</INSDFeature_value>
</INSDFeature>

<INSDFeature>
  <INSDFeature_key>note</INSDFeature_key>
  <INSDFeature_qua>
    <INSDFeature_value>sequence data</INSDFeature_value>
    <INSDFeature_value>standards</INSDFeature_value>
  </INSDFeature_qua>
</INSDFeature>
<table>
<thead>
<tr>
<th>Feature</th>
<th>Source</th>
<th>Location</th>
<th>Type</th>
<th>Organism</th>
<th>Mol Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1..15</td>
<td>source</td>
<td>1..15</td>
<td>DNA</td>
<td>Homo sapiens</td>
<td>genomic DNA</td>
<td>other DNA</td>
</tr>
<tr>
<td>1..10</td>
<td>source</td>
<td>1..10</td>
<td>DNA</td>
<td>synthetic construct</td>
<td>other DNA</td>
<td>other DNA</td>
</tr>
<tr>
<td>1..10</td>
<td>source</td>
<td>1..10</td>
<td>DNA</td>
<td>Homo sapiens</td>
<td>synthetic construct</td>
<td>other DNA</td>
</tr>
</tbody>
</table>

**Sequence Data**

```
ATGCCTGGAC
```

**Sequence ID Number**

```
//ST.26
```
This sequence is one of four branches of a branched polynucleotide. This sequence is one of four branches of a branched polynucleotide.
This sequence is one of four branches of a branched polynucleotide. The sequence is

gcggatttagctcagctgggagagcgccagactgaatanctggagtcctgtgtncgatccacagaattcgcacca

cacataggcatctcctagtgcaggaaga

<table>
<thead>
<tr>
<th>DNA</th>
<th>RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAT</td>
<td>PAT</td>
</tr>
<tr>
<td>cacataggcatctcctagtgcaggaaga</td>
<td>cacataggcatctcctagtgcaggaaga</td>
</tr>
</tbody>
</table>

Ref.: Standards - ST.26
<table>
<thead>
<tr>
<th>Feature</th>
<th>Qualifiers</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td></td>
<td>gatcatacgcactatacgcactatacgcactatacgcactgtac</td>
</tr>
<tr>
<td>ST.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>modified_base</td>
<td>reported base</td>
<td></td>
</tr>
<tr>
<td>misc_difference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Note</td>
<td></td>
<td></td>
</tr>
<tr>
<td>organisms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>organism</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
This sequence is one part of a branched peptide.
This sequence is one part of a branched peptide.

This sequence is one part of a branched peptide.
This sequence is one part of a branched peptide.

ATFGKKKA

FRAGGK

HANDBOOK ON INDUSTRIAL PROPERTY
INFORMATION AND DOCUMENTATION
DGSAKKK/INSDQualifier_value
</INSDQualifier>
</INSDFeature>
</INSDSeq_feature-table>
</INSDSeq>
</SequenceData>

/SequenceData_sequenceIDNumber="33"
<INSDSeq>
<INSDSeq_length>5</INSDSeq_length>
<INSDSeq_moltype>AA</INSDSeq_moltype>
<INSDSeq_division>PA</INSDSeq_division>
<INSDSeq_feature-table>
  <INSDFeature>
    <INSDFeature_key>SOURCE</INSDFeature_key>
    <INSDFeature_location>1..5</INSDFeature_location>
    <INSDFeature_quals>
      <INSDQualifier>
        <INSDQualifier_name>ORGANISM</INSDQualifier_name>
        <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
      </INSDQualifier>
      <INSDQualifier>
        <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
        <INSDQualifier_value>protein</INSDQualifier_value>
      </INSDQualifier>
    </INSDFeature_quals>
  </INSDFeature>
  <INSDFeature>
    <INSDFeature_key>SITE</INSDFeature_key>
    <INSDFeature_location>5</INSDFeature_location>
    <INSDFeature_quals>
      <INSDQualifier>
        <INSDQualifier_name>NOTE</INSDQualifier_name>
        <INSDQualifier_value>Bound to Lys at position 5 of protein</INSDQualifier_value>
      </INSDQualifier>
    </INSDFeature_quals>
  </INSDFeature>
</INSDSeq>
</SequenceData>

/SequenceData_sequenceIDNumber="34"
<INSDSeq>
<INSDSeq_length>13</INSDSeq_length>
<INSDSeq_moltype>DNA</INSDSeq_moltype>
</INSDSeq>
</SequenceData>
<INSDSeq_feature-tables>
  <INSDFeature>
    <INSDFeature_key>source</INSDFeature_key>
    <INSDFeature_location_q>1..58</INSDFeature_location_q>
    <INSDQualifier name="organism">Homo sapiens</INSDQualifier>
    <INSDQualifier name="mol_type">genomic DNA</INSDQualifier>
  </INSDFeature>
  <INSDSeq_sequence>atgcgaattgcc</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<INSDSeq_sequence>ggcgaattgcgat</INSDSeq_sequence>
</SequenceData>
<INSDSeq_sequence>ggcgaattgcgat</INSDSeq_sequence>
</SequenceData>
<INSDSeq_sequence>ggcgaattgcgat</INSDSeq_sequence>
</SequenceData>
<INSDSeq_sequence>taggccatgactaaggctccccatgtgactaaggcgactagcattgactaaggcaagc</INSDSeq_sequence>
</SequenceData>
<INSDSeq_sequence>taggccatgactaaggctccccatgtgactaaggcgactagcattgactaaggcaagc</INSDSeq_sequence>
</SequenceData>
<INSDSeq_sequence>taggccatgactaaggctccccatgtgactaaggcgactagcattgactaaggcaagc</INSDSeq_sequence>
</SequenceData>
<INSDSeq_sequence>taggccatgactaaggctccccatgtgactaaggcgactagcattgactaaggcaagc</INSDSeq_sequence>
</SequenceData>
<INSDSeq_sequence>taggccatgactaaggctccccatgtgactaaggcgactagcattgactaaggcaagc</INSDSeq_sequence>
</SequenceData>
The linker which is attached to the 5' end to a thymidine is attached at its 3' oxygen of the thymidine. The linker is (4-(3-hydroxybenzamido)butyl) phosphinic acid which is attached to the 5' end to a thymidine.
Up to 98 copies of GGGX can be deleted.
Ref.: Standards - ST.26

Date: September 2017
<INSDSeq>
  <INSDSeq_feature-tables>
    <INSDSeq_sequence>MTNITDGlAVLqGRDfNP56HfYLAR4CW4</INSDSeq_sequence>
  </INSDSeq_feature-tables>
  <INSDSeq_sequence>INSDSeq_feature INSDSeq_moltype INSDSeq_length INSDSeq_sequence</INSDSeq_sequence>
</INSDSeq>

<INSDSeq>
  <INSDSeq_feature-tables>
    <INSDSeq_sequence>MLRQTNEVA</INSDSeq_sequence>
  </INSDSeq_feature-tables>
  <INSDSeq_sequence>INSDSeq_feature INSDSeq_moltype INSDSeq_length INSDSeq_sequence</INSDSeq_sequence>
</INSDSeq>

<INSDSeq>
  <INSDSeq_feature-tables>
    <INSDSeq_sequence>PAT</INSDSeq_sequence>
  </INSDSeq_feature-tables>
  <INSDSeq_sequence>INSDSeq_feature INSDSeq_moltype INSDSeq_length INSDSeq_sequence</INSDSeq_sequence>
</INSDSeq>

<INSDSeq>
  <INSDSeq_feature-tables>
    <INSDSeq_sequence>1..10</INSDSeq_sequence>
  </INSDSeq_feature-tables>
  <INSDSeq_sequence>INSDSeq_feature INSDSeq_moltype INSDSeq_length INSDSeq_sequence</INSDSeq_sequence>
</INSDSeq>

<INSDSeq>
  <INSDSeq_feature-tables>
    <INSDSeq_sequence>1..10</INSDSeq_sequence>
  </INSDSeq_feature-tables>
  <INSDSeq_sequence>INSDSeq_feature INSDSeq_moltype INSDSeq_length INSDSeq_sequence</INSDSeq_sequence>
</INSDSeq>

<INSDSeq>
  <INSDSeq_feature-tables>
    <INSDSeq_sequence>1..10</INSDSeq_sequence>
  </INSDSeq_feature-tables>
  <INSDSeq_sequence>INSDSeq_feature INSDSeq_moltype INSDSeq_length INSDSeq_sequence</INSDSeq_sequence>
</INSDSeq>

<INSDSeq>
  <INSDSeq_feature-tables>
    <INSDSeq_sequence>1..10</INSDSeq_sequence>
  </INSDSeq_feature-tables>
  <INSDSeq_sequence>INSDSeq_feature INSDSeq_moltype INSDSeq_length INSDSeq_sequence</INSDSeq_sequence>
</INSDSeq>

<INSDSeq>
  <INSDSeq_feature-tables>
    <INSDSeq_sequence>1..10</INSDSeq_sequence>
  </INSDSeq_feature-tables>
  <INSDSeq_sequence>INSDSeq_feature INSDSeq_moltype INSDSeq_length INSDSeq_sequence</INSDSeq_sequence>
</INSDSeq>

<INSDSeq>
  <INSDSeq_feature-tables>
    <INSDSeq_sequence>1..10</INSDSeq_sequence>
  </INSDSeq_feature-tables>
  <INSDSeq_sequence>INSDSeq_feature INSDSeq_moltype INSDSeq_length INSDSeq_sequence</INSDSeq_sequence>
</INSDSeq>

<INSDSeq>
  <INSDSeq_feature-tables>
    <INSDSeq_sequence>1..10</INSDSeq_sequence>
  </INSDSeq_feature-tables>
  <INSDSeq_sequence>INSDSeq_feature INSDSeq_moltype INSDSeq_length INSDSeq_sequence</INSDSeq_sequence>
</INSDSeq>

<INSDSeq>
  <INSDSeq_feature-tables>
    <INSDSeq_sequence>1..10</INSDSeq_sequence>
  </INSDSeq_feature-tables>
  <INSDSeq_sequence>INSDSeq_feature INSDSeq_moltype INSDSeq_length INSDSeq_sequence</INSDSeq_sequence>
</INSDSeq>
Ref.: Standards - ST.26

Date: September 2017
<INSDSeq_sequence>
KXXKXXXXXXXXXXXXXXXXXXXXXXXXXKXXXKXXXXXXXXXXXXXXXXXXXXXXXXXKXXK
</INSDSeq_sequence>

<INSDSeq_moltype>protein</INSDSeq_moltype>

<INSDQualifier name="MOL_TYPE">

</INSDQualifier>

<INSDSeq_division>PAT</INSDSeq_division>

<INSDFeature key="SOURCE">

</INSDFeature>

<INSDQualifier name="ORGANISM">

</INSDQualifier>

<INSDQualifier name="MOL_TYPE">

</INSDQualifier>

<INSDQualifier name="protein">

</INSDQualifier>

<INSDQualifier name="VARIANT">

</INSDQualifier>

<INSDQualifier name="NOTE">

</INSDQualifier>

<INSDQualifier name="Up to 10 residues may be deleted">

</INSDQualifier>

<INSDQualifier name="Up to 10 X residues may be deleted">

</INSDQualifier>

<INSDQualifier name="Up to 10 X residues may be inserted">

</INSDQualifier>

<INSDQualifier name="Up to 10 X residues may be deleted">

</INSDQualifier>

<INSDQualifier name="Up to 10 X residues may be inserted">

</INSDQualifier>

<INSDQualifier name="Up to 10 residues may be deleted">

</INSDQualifier>

<INSDQualifier name="Up to 10 X residues may be inserted">

</INSDQualifier>

<INSDQualifier name="Up to 10 residues may be deleted">

</INSDQualifier>

<INSDQualifier name="Up to 10 X residues may be inserted">

</INSDQualifier>

<INSDQualifier name="Up to 10 residues may be deleted">

</INSDQualifier>

<INSDQualifier name="Up to 10 X residues may be inserted">

</INSDQualifier>

<INSDQualifier name="Up to 10 residues may be deleted">

</INSDQualifier>

<INSDQualifier name="Up to 10 X residues may be inserted">

</INSDQualifier>

<INSDQualifier name="Up to 10 residues may be deleted">

</INSDQualifier>

<INSDQualifier name="Up to 10 X residues may be inserted">

</INSDQualifier>
<INSDSeq>
  <SequenceData_sequence1Number="57"/>
  <INSDSeq_length>45</INSDSeq_length>
  <INSDSeq_moltype>AA</INSDSeq_moltype>
  <INSDSeq_division>PAT</INSDSeq_division>
  <INSDSeq_feature-table>
    <INSDFeature>
      <INSDFeature_key>SOURCE</INSDFeature_key>
      <INSDFeature_location>1..45</INSDFeature_location>
      <INSDFeature_qual>
        <INSDQualifier name="ORGANISM">Homo sapiens</INSDQualifier>
      </INSDFeature_qual>
    </INSDFeature>
    <INSDFeature>
      <INSDFeature_key>SITE</INSDFeature_key>
      <INSDFeature_location>1..45</INSDFeature_location>
      <INSDFeature_qual>
        <INSDQualifier name="NOTE">This residue is linked N-terminally to a peptide having an N-terminal Gly-Gly and a gap of undefined length</INSDQualifier>
      </INSDFeature_qual>
    </INSDFeature>
  </INSDSeq_feature-table>
  <INSDSeq_sequence>atgaccggaaataaacctgaaaccgatgtttacgaaattttatga</INSDSeq_sequence>
</INSDSeq>

<INSDSeq>
  <SequenceData_sequence1Number="58"/>
  <INSDSeq_length>58</INSDSeq_length>
  <INSDSeq_moltype>DNA</INSDSeq_moltype>
  <INSDSeq_division>PAT</INSDSeq_division>
  <INSDSeq_feature-table>
    <INSDFeature>
      <INSDFeature_key>site</INSDFeature_key>
      <INSDFeature_location>1..45</INSDFeature_location>
      <INSDFeature_qual>
        <INSDQualifier name="organism">Homo sapiens</INSDQualifier>
        <INSDQualifier name="mol_type">genomic DNA</INSDQualifier>
      </INSDFeature_qual>
    </INSDFeature>
    <INSDFeature>
      <INSDFeature_key>translation</INSDFeature_key>
      <INSDFeature_location>1..45</INSDFeature_location>
      <INSDFeature_qual>
        <INSDQualifier name="protein_id">MTONKFETDIYEL</INSDQualifier>
      </INSDFeature_qual>
    </INSDFeature>
    <INSDFeature>
      <INSDFeature_key>protein</INSDFeature_key>
      <INSDFeature_location>1..45</INSDFeature_location>
      <INSDFeature_qual>
        <INSDQualifier name="protein">MTONKFETDIYEL</INSDQualifier>
      </INSDFeature_qual>
    </INSDFeature>
  </INSDSeq_feature-table>
  <INSDSeq_sequence>atgaccggaaataaacctgaaaccgatgtttacgaaattttatga</INSDSeq_sequence>
</INSDSeq>

<SequenceData sequenceIDNumber="59">
  <INSDSeg>
    <INSDSeq_length>14</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>ST</INSDSeq_division>
    <INSDSeq_feature>table</INSDSeq_feature>
    <INSDFeature key>source</INSDFeature key>
    <INSDFeature_location>1..14</INSDFeature_location>
    <INSDFeature_qual>
      <INSDFeature_value>
        <INSDFeature qualifier>organism</INSDFeature qualifier>
        <INSDFeature_value> Homo sapiens </INSDFeature_value>
      </INSDFeature_value>
      <INSDFeature_value>
        <INSDFeature qualifier>MOL_TYPE</INSDFeature qualifier>
        <INSDFeature_value>DNA</INSDFeature_value>
      </INSDFeature_value>
    </INSDFeature_qual>
  </INSDSeg>
</SequenceData>

<SequenceData sequenceIDNumber="60">
  <INSDSeg>
    <INSDSeq_length>22</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>ST</INSDSeq_division>
    <INSDSeq_feature>table</INSDSeq_feature>
    <INSDFeature key>source</INSDFeature key>
    <INSDFeature_location>1..22</INSDFeature_location>
    <INSDFeature_qual>
      <INSDFeature_value>
        <INSDFeature qualifier>organism</INSDFeature qualifier>
        <INSDFeature_value> Drosophila melanogaster </INSDFeature_value>
      </INSDFeature_value>
      <INSDFeature_value>
        <INSDFeature qualifier>MOL_TYPE</INSDFeature qualifier>
        <INSDFeature_value>DNA</INSDFeature_value>
      </INSDFeature_value>
    </INSDFeature_qual>
  </INSDSeg>
</SequenceData>

<SequenceData sequenceIDNumber="61">
  <INSDSeg>
    <INSDSeq_length>19</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>ST</INSDSeq_division>
    <INSDSeq_feature>table</INSDSeq_feature>
    <INSDFeature key>source</INSDFeature key>
    <INSDFeature_location>1..19</INSDFeature_location>
    <INSDFeature_qual>
      <INSDFeature_value>
        <INSDFeature qualifier>organism</INSDFeature qualifier>
        <INSDFeature_value> Drosophila virilis </INSDFeature_value>
      </INSDFeature_value>
      <INSDFeature_value>
        <INSDFeature qualifier>MOL_TYPE</INSDFeature qualifier>
        <INSDFeature_value>DNA</INSDFeature_value>
      </INSDFeature_value>
    </INSDFeature_qual>
  </INSDSeg>
</SequenceData>

<SequenceData sequenceIDNumber="62">
  <INSDSeg>
    <INSDSeq_length>22</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>ST</INSDSeq_division>
    <INSDSeq_feature>table</INSDSeq_feature>
    <INSDFeature key>source</INSDFeature key>
    <INSDFeature_location>1..22</INSDFeature_location>
    <INSDFeature_qual>
      <INSDFeature_value>
        <INSDFeature qualifier>organism</INSDFeature qualifier>
        <INSDFeature_value> Homo sapiens </INSDFeature_value>
      </INSDFeature_value>
      <INSDFeature_value>
        <INSDFeature qualifier>MOL_TYPE</INSDFeature qualifier>
        <INSDFeature_value>DNA</INSDFeature_value>
      </INSDFeature_value>
    </INSDFeature_qual>
  </INSDSeg>
</SequenceData>
<table>
<thead>
<tr>
<th>SequenceData</th>
<th>sequenceIDNumber</th>
<th>INSDSeq_sequence</th>
<th>source</th>
<th>PAT</th>
<th>Venue</th>
<th>mol_type</th>
<th>organism</th>
<th>MOL_TYPE</th>
<th>ORGANISM</th>
<th>MOL_TYPE</th>
<th>ORGANISM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>tggagctgtatgatcttt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Homo sapiens</td>
<td>protein</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Drosophila simulans</td>
<td>protein</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>protein</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**HANDBOOK ON INDUSTRIAL PROPERTY INFORMATION AND DOCUMENTATION**

Ref.: Standards - ST.26  
Date: September 2017
<INSDFeature>
  <INSDFeature_key>VARIANT</INSDFeature_key>
  <INSDFeature_location>1..7</INSDFeature_location>
  <INSDFeature_qua>
    <INSDFeature.qual>
      <INSDFeature.qualifier>NOTE</INSDFeature.qualifier>
      <INSDFeature.qualifier.value>replace with I</INSDFeature.qualifier.value>
    </INSDFeature.qual>
  </INSDFeature_qua>
</INSDFeature>

<INSDFeature>
  <INSDFeature_key>VARIANT</INSDFeature_key>
  <INSDFeature_location>7</INSDFeature_location>
  <INSDFeature_qua>
    <INSDFeature.qual>
      <INSDFeature.qualifier>NOTE</INSDFeature.qualifier>
      <INSDFeature.qualifier.value>replace with F</INSDFeature.qualifier.value>
    </INSDFeature.qual>
  </INSDFeature_qua>
</INSDFeature>

<INSDFeature>
  <INSDFeature_key>SOURCE</INSDFeature_key>
  <INSDFeature.location>1..7</INSDFeature.location>
  <INSDFeature_qua>
    <INSDFeature.qual>
      <INSDFeature.qualifier>ORGANISM</INSDFeature.qualifier>
      <INSDFeature.qualifier.value>Homo sapiens</INSDFeature.qualifier.value>
    </INSDFeature.qual>
  </INSDFeature_qua>
</INSDFeature>

<INSDFeature>
  <INSDFeature_key>MOL_TYPE</INSDFeature_key>
  <INSDFeature.sequencevalue>protein</INSDFeature.sequencevalue>
</INSDFeature>

<INSDFeature>
  <INSDFeature_key>SOURCE</INSDFeature_key>
  <INSDFeature.location>7</INSDFeature.location>
  <INSDFeature_qua>
    <INSDFeature.qual>
      <INSDFeature.qualifier>MOL_TYPE</INSDFeature.qualifier>
      <INSDFeature.qualifier.value>protein</INSDFeature.qualifier.value>
    </INSDFeature.qual>
  </INSDFeature_qua>
</INSDFeature>

<INSDFeature>
  <INSDFeature_key>NOTE</INSDFeature_key>
  <INSDFeature_location>7</INSDFeature_location>
  <INSDFeature_qua>
    <INSDFeature.qual>
      <INSDFeature.qualifier>T</INSDFeature.qualifier>
      <INSDFeature.qualifier_value>replace with T</INSDFeature.qualifier_value>
    </INSDFeature.qual>
  </INSDFeature_qua>
</INSDFeature>

<INSDFeature>
  <INSDFeature_key>VARIANT</INSDFeature_key>
  <INSDFeature_location>7</INSDFeature_location>
  <INSDFeature_qua>
    <INSDFeature.qual>
      <INSDFeature.qualifier>MOL_TYPE</INSDFeature.qualifier>
      <INSDFeature.qualifier.value>protein</INSDFeature.qualifier.value>
    </INSDFeature.qual>
  </INSDFeature_qua>
</INSDFeature>
<INSSeq_sequence>MKTFAALLSAVTLALSVRQAQAAMVWQCGGTFOWTGETTCVAIGSVCTLSLSSYQCVPOATSSAPAAPSATTSGPAPTDGTCSASGAWPFLN</INSSeq_sequence>
<table>
<thead>
<tr>
<th>ST.26</th>
<th>INSDSeq_length</th>
<th>INSDSeq_moltype</th>
<th>INSDSeq_division</th>
<th>INSDSeq_moltype</th>
<th>INSDSeq_length</th>
<th>INSDSeq_moltype</th>
<th>INSDSeq_division</th>
<th>INSDSeq_moltype</th>
<th>INSDSeq_length</th>
<th>INSDSeq_moltype</th>
<th>INSDSeq_division</th>
<th>INSDSeq_moltype</th>
</tr>
</thead>
<tbody>
<tr>
<td>1..32</td>
<td>81</td>
<td>PAT</td>
<td>Homo sapiens</td>
<td>protein</td>
<td>80</td>
<td>PAT</td>
<td>Homo sapiens</td>
<td>protein</td>
<td>9</td>
<td>PAT</td>
<td>Homo sapiens</td>
<td>protein</td>
</tr>
<tr>
<td>1..8</td>
<td>79</td>
<td>PAT</td>
<td>Homo sapiens</td>
<td>protein</td>
<td>8</td>
<td>PAT</td>
<td>Homo sapiens</td>
<td>protein</td>
<td>9</td>
<td>PAT</td>
<td>Homo sapiens</td>
<td>protein</td>
</tr>
<tr>
<td>1..9</td>
<td>79</td>
<td>PAT</td>
<td>Homo sapiens</td>
<td>protein</td>
<td>8</td>
<td>PAT</td>
<td>Homo sapiens</td>
<td>protein</td>
<td>9</td>
<td>PAT</td>
<td>Homo sapiens</td>
<td>protein</td>
</tr>
</tbody>
</table>
<INSDQualifier><INSDQualifier name="NOTE"></INSDQualifier></INSDQualifier>
<INSDQualifier><INSDQualifier name="M or L"></INSDQualifier></INSDQualifier>
<INSDFeature><INSDFeature key="SOURCE"></INSDFeature></INSDFeature>
<INSDSeq_sequence>LEGNEQPINAASKIKRRHPQYDRKTVNHDDILK</INSDSeq_sequence></INSDSeq>
</SequenceData><SequenceData_sequenceIDNumber="82">
<INSDSeq><INSDSeq_length>32</INSDSeq_length></INSDSeq>
<INSDSeq_moltype>AA</INSDSeq_moltype>
<INSDSeq_division>PAT</INSDSeq_division>
<INSDSeq_feature tabela>
<INSDFeature><INSDFeature_key="SOURCE"></INSDFeature_key></INSDFeature>
<INSDFeature_location>1..32</INSDFeature_location></INSDFeature>
<INSDFeature><INSDFeature_key="ORGANISM"></INSDFeature_key></INSDFeature>
<INSDFeature><INSDFeature_value="Homo sapiens"></INSDFeature_value></INSDFeature>
<INSDFeature><INSDFeature_name="MOL_TYPE"></INSDFeature_name></INSDFeature>
<INSDFeature><INSDFeature_value="protein"></INSDFeature_value></INSDFeature>
</INSDSeqFeature>
</INSDSeq>
</SequenceData><SequenceData_sequenceIDNumber="83">
<INSDSeq><INSDSeq_length>32</INSDSeq_length></INSDSeq>
<INSDSeq_moltype>AA</INSDSeq_moltype>
<INSDSeq_division>PAT</INSDSeq_division>
<INSDSeq_feature tabela>
<INSDFeature><INSDFeature_key="SOURCE"></INSDFeature_key></INSDFeature>
<INSDFeature_location>1..32</INSDFeature_location></INSDFeature>
<INSDFeature><INSDFeature_key="ORGANISM"></INSDFeature_key></INSDFeature>
<INSDFeature><INSDFeature_value="Pongo abelii"></INSDFeature_value></INSDFeature>
<INSDFeature><INSDFeature_name="MOL_TYPE"></INSDFeature_name></INSDFeature>
<INSDFeature><INSDFeature_value="protein"></INSDFeature_value></INSDFeature>
</INSDSeqFeature>
</INSDSeq>
</SequenceData><SequenceData_sequenceIDNumber="84">
<INSDSeq><INSDSeq_length>32</INSDSeq_length></INSDSeq>
<INSDSeq_moltype>AA</INSDSeq_moltype>
<INSDSeq_division>PAT</INSDSeq_division>
<INSDSeq_feature tabela>
<INSDFeature><INSDFeature_key="SOURCE"></INSDFeature_key></INSDFeature>
<INSDFeature_location>1..32</INSDFeature_location></INSDFeature>
<INSDFeature><INSDFeature_key="ORGANISM"></INSDFeature_key></INSDFeature>
<INSDFeature><INSDFeature_value="Papio anubis"></INSDFeature_value></INSDFeature>
<INSDFeature><INSDFeature_name="MOL_TYPE"></INSDFeature_name></INSDFeature>
<INSDFeature><INSDFeature_value="protein"></INSDFeature_value></INSDFeature>
</INSDSeqFeature>
</INSDSeq>
</SequenceData>
[End of Annex VI and end of Standard]