1. The Committee on WIPO Standards (CWS), at its first session held in October 2010, agreed to create Task No. 44 for the preparation of recommendations on the presentation of nucleotide and amino acid sequence listings based on eXtensible Markup Language (XML) for adoption as a WIPO standard. The CWS also decided to establish a task force to handle this Task (SEQL Task Force). The European Patent Office (EPO) was designated as the Task Force Leader. (See paragraphs 27 to 30 of document CWS/1/10, Task No. 44 in document CWS/3/12, the description of the Task is also provided in Annex I to the present document.)

2. Following the above decision by the CWS, representatives of 13 industrial property offices (IPO) and the International Bureau were nominated to participate in the Task Force. At its second and third sessions, the CWS noted the information that the EPO, in its capacity of the Task Force Leader, provided on the status of the SEQL Task Force discussion, including the work plan for preparation of the recommendations. (See documents CWS/2/5 and CWS/3/6.)

3. After the third session of the CWS, the SEQL Task Force continued its discussion in the Wiki forum. The report prepared by the Task Force Leader on the work carried out by the Task Force is reproduced in Annex I to the present document.

4. Following the above-mentioned request by the CWS, the SEQL Task Force has prepared a proposal of a new standard for consideration and approval by the CWS. The proposed name of the new standard is "WIPO Standard ST.26 – Recommended standard for the presentation of nucleotide and amino acid sequence listings using XML (eXtensible Markup Language)". The draft new WIPO Standard ST.26, which contains the main body and five annexes, is reproduced in Annex II to the present document.
5. The SEQL Task Force was also requested by the CWS to liaise with the appropriate PCT body with regard to the possible impact of new Standard ST.26 on Annex C to the Administrative Instructions under the PCT (see paragraph 29(c) of document CWS/1/10). Provisions related to the transition from WIPO Standard ST.25 to the new WIPO Standard ST.26 are currently under discussion by the Task Force members. It is planned to present them for consideration by the CWS at its next session to be held in 2015 (see paragraph 10 “Roadmap” in Annex I to the present document).

6. IPOs are requested to postpone the preparations for implementation of the new WIPO Standard ST.26 until said provisions on the transition are approved by the CWS. Meanwhile, Standard ST.25 should continue to be used. Having this in mind, on the condition that the new standard is adopted at the present (fourth) session of the CWS, the Task Force proposes to include the following Editorial Note to the new standard:

“Editorial Note by the International Bureau

“The 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 fifth session to be held in 2015. Meanwhile, Standard ST.25 should continue to be used.

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

“The Committee on WIPO Standards (CWS) adopted the present standard at [its fourth session on May 16, 2014]."

7. The CWS is invited to:

(a) note the status report on the work of SEQL Task Force provided in Annex I to the present document;

(b) adopt “WIPO Standard ST.26 – Recommended standard for the presentation of nucleotide and amino acid sequence listings using XML (eXtensible Markup Language)” as the name of the proposed standard;

(c) consider and adopt WIPO Standard ST.26, as reproduced in the Annex II to the present document;

(d) consider and approve the Editorial Note to be included in WIPO Standard ST.26 (see paragraph 6, above); and

(e) request the SEQL Task Force to prepare a proposal for the transition provisions, as referred to in paragraph 5, above, and present them for consideration and approval by the CWS at its fifth session.

[Annexes follow]
REPORT ON THE PREPARATION OF A NEW WIPO STANDARD ON THE PRESENTATION OF NUCLEOTIDE AND AMINO ACID SEQUENCE LISTINGS USING EXTENSIBLE MARKUP LANGUAGE (XML)

Document prepared by the European Patent Office (EPO)

BACKGROUND

1. The Task Force on Sequence Listings was created by the Committee on WIPO Standards (CWS), at its first session (October 25 to 29, 2010), to deal with Task No. 44 (see paragraph 29 of document CWS/1/10):

   “Prepare a recommendation on the presentation of nucleotide and amino acid sequence listings based on eXtensible Markup Language (XML) for adoption as a WIPO standard. The proposal of the new WIPO standard should be presented along with a report on the impact of the said standard on the current WIPO Standard ST.25, including the proposed necessary changes to Standard ST.25.”

2. The Task Force was also requested:

   “To liaise with the appropriate PCT body with regard to the possible impact of such standard on Annex C to the Administrative Instructions under the PCT.”

3. The European Patent Office (EPO) was assigned the role of Task Force Leader and has since then held six rounds of discussions on WIPO’s Wiki and submitted a final draft for public consultation. The principle of differentiating the technical aspects of ST.25 from Annex C (PCT Administrative Instructions) was agreed upon at the eighteenth session Meeting of International Authorities in February 2011 (see paragraphs 88 to 92 of document PCT/MIA/18/16) and at the fourth session of the PCT Working Group in June 2011 (see paragraphs 180 to 188 of document PCT/WG/4/17).

4. On the basis of the comments received from the Task Force members, a final round of discussions was held in order to achieve a common agreement on the Standard requirements.

PROGRESS REPORT

5. The Task Force started operating in February 2011 on the basis of drafts prepared by the EPO. Many Offices participated in the process and posted useful comments on WIPO’s related Wiki.

6. In March 2012, the Task Force finalized a draft of the standard that could be used by the offices to consult their respective public. Several significant issues were raised by public comments and they have been tackled in cooperation with the database providers DDBJ, EBI, NCBI.

7. The sixth round of discussions was finished on September 2013 and the draft incorporating the improvements resulting from the public consultation and further discussions among the Task Force members and Database providers, was posted on WIPO’s wiki for a final review.
8. On the basis of the comments received from the Task Force members, a final round of discussions took place aimed at achieving a common agreement on the Standard requirements. On a provisional basis, the Task Force named this Standard ST.26. The main body and annexes proposed by the Task Force for consideration and approval of the CWS introduce the following improvements when compared to the current ST.25:

(a) All (PCT) procedural issues are transferred to the PCT Administrative Instructions: the new standard is focused on technical aspects, thus to enable an optimal presentation of the sequence listings (the biotech-related part) and the appropriate format of the submission (namely XML);

(b) The biotech-related part has been considerably improved to reflect modern industry standards, for example:
    - inclusion of modified nucleotides and amino acids not previously provided for (e.g. D-amino acids, PNA, morpholinos etc.) which have gained importance in industry and need to be electronically searchable;
    - clear instructions for gapped sequences and sequence variants;
    - clarification with regard to features and annotations;
    - consistency with latest public biological sequence repositories consortia requirements (INSDC and UniProt); and
    - the XML definition is self-contained and independent of ST.36 or ST.96.

(c) The syntax provided by the Document Type Definition (DTD) used in the ST.26 increases the data accuracy and enables automatic data quality control.

9. The Task Force will continue to work on the transition aspects in 2014 and 2015 with the objective to submit for the consideration and approval of the CWS, at its fifth session, the recommendations for the transition from ST.25 to ST.26.

ROADMAP

10. A new round of discussions will continue after the CWS/4 session, focused on the preparation of the recommendations for the transition phase to be presented at the CWS session in 2015.

[Annex II follows]
TABLE OF CONTENTS

INTRODUCTION ........................................................................................................................................................... 2
DEFINITIONS ................................................................................................................................................................ 2
SCOPE .......................................................................................................................................................................... 3
REFERENCES .............................................................................................................................................................. 3
PRESENTATION OF SEQUENCES .............................................................................................................................. 3
  Nucleotide sequences ................................................................................................................................................ 3
  Amino acid sequences ............................................................................................................................................... 5
  Presentation of special situations ............................................................................................................................... 7
STRUCTURE OF THE SEQUENCE LISTING IN XML .................................................................................................. 7
  Root element .............................................................................................................................................................. 8
  General information part ............................................................................................................................................. 8
  Sequence data part .................................................................................................................................................. 12
  Feature table ............................................................................................................................................................ 14
  Feature keys ............................................................................................................................................................ 14
  Mandatory feature keys ............................................................................................................................................ 14
  Feature location ....................................................................................................................................................... 14
  Feature qualifiers ...................................................................................................................................................... 16
  Mandatory feature qualifiers ..................................................................................................................................... 16
  Qualifier elements .................................................................................................................................................... 16
  Free text ................................................................................................................................................................... 18
  Coding sequences .................................................................................................................................................... 18
  Variants .................................................................................................................................................................... 19

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)
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, 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 as modified according to paragraph 29.

   (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) “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.

   (d) “nucleotide” means any nucleotide or nucleotide analog that can be represented using any of the symbols set forth in Annex I (see Section 1, Table 1). Nucleotides may contain, inter alia, a modified or synthetic purine or pyrimidine base, or a modified or synthetic ribose or deoxyribose, and may be joined by a modified or synthetic 3’ to 5’ inter-nucleoside linkage, i.e., any chemical moiety that provides the same structural function as the phosphate moiety of DNA or RNA, such as a phosphorothioate moiety.

   (e) “residue” means any individual nucleotide or amino acid in a sequence.

   (f) “sequence identification number” means a unique number (integer) assigned to each sequence in the sequence listing.

   (g) “sequence listing” means a part of the description of the patent application as filed or a document filed subsequently to the application, which presents the disclosed nucleotide and/or amino acid sequence(s), along with any further description.

   (h) “specifically defined” means 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.

   (i) “unknown” nucleotide or amino acid means that a single nucleotide or amino acid is present but its identity is unknown or not disclosed.
SCOPE

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

5. 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.

6. 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 is:

   (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
   (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.

7. A sequence listing must not include any sequences having fewer than ten specifically defined nucleotides, or fewer than four specifically defined amino acids.

REFERENCES

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

   - International Nucleotide Sequence Database Collaboration (INSDC)  http://www.insdc.org/;
   - ISO 639-1 - Codes for the representation of names of languages  Part 1: Alpha-2 code;
   - 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

9. Each sequence must be assigned a separate sequence identification number. 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

10. A nucleotide sequence must be presented only by a single strand, in the 5'-end to 3'-end direction from left to right. The designations 5' and 3' must not be present 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.

11. Numbering of the nucleotide positions must start at the first base of the sequence with number 1. It must be continuous through the whole sequence in the direction 5' to 3'.

12. The above numbering method for nucleotide sequences is also applicable to nucleotide sequences that are circular in configuration. In this case, the applicant must choose the nucleotide with which numbering begins.

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 by paragraph 18.

15. Where an ambiguity symbol (representing two or more alternative bases) 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 any one of “a”, “c”, “g”, or “tu” except where it is used with a further description as provided by paragraphs 16 and 17 or 20. The symbol “n” may not be used to represent anything other than a nucleotide. A single modified or “unknown” nucleotide may be represented by the symbol “n”, together with a further description, as provided in paragraphs 16 and 17 or 20.

16. Modified nucleotides should be represented in the sequence as the corresponding unmodified bases, i.e., “a”, “c”, “g” or “t” whenever possible. Any modified nucleotide in a sequence that cannot otherwise be represented by any other symbol in Annex I (see Section 1, Table 1), such as non-naturally occurring nucleotides, must be represented by the symbol “n”. Where the symbol “n” is used to represent a modified nucleotide it is the equivalent of only one residue.

17. A modified nucleotide must be further described in the feature table (see paragraph 59 et seq.) using the feature key “modified_base” and the mandatory qualifier “mod_base” in conjunction with a single abbreviation from Annex I (see Section 2, Table 2) as the qualifier value; if the abbreviation is “OTHER”, the complete unabbreviated name of the modified base must be provided as the value in a “note” qualifier. The abbreviations (or full names) provided in Annex I (see Section 2, Table 2) referred to above must not be used in the sequence itself.

18. Uracil in DNA or thymine in RNA are considered modified nucleotides and must be represented in the sequence as “t” and be further described in the feature table using the feature key “modified_base”, the qualifier “mod_base” with “OTHER” as the qualifier value and the qualifier “note” with “uracil” or “thymine”, respectively, as the qualifier value.

19. The following examples illustrate the presentation of modified nucleotides according to paragraphs 16 and 17 above:

Example 1: Modified nucleotide using an abbreviation from Annex I (see Section 2, Table 2)

```xml
<INSDFeature>
  <INSDFeature_key>modified_base</INSDFeature_key>
  <INSDFeature_location>15</INSDFeature_location>
  <INSDFeature_quals>
    <INSDQualifier>
      <INSDQualifier_name>mod_base</INSDQualifier_name>
      <INSDQualifier_value>i</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>
```

Example 2: Modified nucleotide using “OTHER” from Annex I (see Section 2, Table 2)

```xml
<INSDFeature>
  <INSDFeature_key>modified_base</INSDFeature_key>
  <INSDFeature_location>4</INSDFeature_location>
  <INSDFeature_quals>
    <INSDQualifier>
      <INSDQualifier_name>mod_base</INSDQualifier_name>
      <INSDQualifier_value>OTHER</INSDQualifier_value>
    </INSDQualifier>
    <INSDQualifier>
      <INSDQualifier_name>note</INSDQualifier_name>
      <INSDQualifier_value>xanthine</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>
```

20. 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.
21. 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 the syntax "x..y" as the location descriptor in the element INSDFeature_location (see paragraphs 65 to 72). For presentation of sequence variants, i.e., deletions, insertions or substitutions, see paragraphs 92 to 97.

22. The following example illustrates the presentation of a region of modified nucleotides for which the same description applies, according to paragraph 21 above:

```xml
<INSDFeature>
  <INSDFeature_key>modified_base</INSDFeature_key>
  <INSDFeature_location>358..485</INSDFeature_location>
  <INSDFeature_quals>
    <INSDQualifier>
      <INSDQualifier_name>mod_base</INSDQualifier_name>
      <INSDQualifier_value>OTHER</INSDQualifier_value>
    </INSDQualifier>
    <INSDQualifier>
      <INSDQualifier_name>note</INSDQualifier_name>
      <INSDQualifier_value>isoguanine</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>
```

Amino acid sequences

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

24. Numbering of amino acid positions must start at the first amino acid of the sequence, with number 1, including amino acids preceding the mature protein, for example, pre-sequences, pro-sequences, pre-pro-sequences and signal sequences. It must be contiguous through the whole sequence in the amino to carboxy direction.

25. 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.

26. 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", "S", "T", "W", "Y", or "V", except where it is used with a further description as provided by paragraphs 28 to 30 or 31 to 33. The symbol "X" may not be used to represent anything other than an amino acid. A single amino acid may be represented by the symbol "X", together with a further description, as provided in paragraphs 28 to 30 or 31 to 33. For presentation of sequence variants, i.e., deletions, insertions, or substitutions, see paragraphs 92 to 97.

27. Amino acid sequences separated by one or more blank spaces or internal terminator symbols, for example, "Ter" or asterisk "*" or period ".", in a disclosure, 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 6. Each such separate sequence must be presented in the sequence listing with its own sequence identification number, using only the symbols set forth in Annex I (see Section 3, Table 3). Terminator symbols and spaces must not be used in sequences in a sequence listing.

28. 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 other symbol in Annex I (see Section 3, Table 3), must be represented by "X". The symbol "X" is the equivalent of only one residue.

29. A modified amino acid must be further described in a feature table (see paragraph 60 et seq.). The feature key "MOD_RES" must be used for post-translationally modified amino acids together with the qualifier “NOTE” and the feature key “SITE” for other modified amino acids together with the qualifier “NOTE”. 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.

30. The following examples illustrate the presentation of modified amino acids according to paragraph 29 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>mod_base</INSDQualifier_name>
      <INSDQualifier_value>OTHER</INSDQualifier_value>
    </INSDQualifier>
    <INSDQualifier>
      <INSDQualifier_name>note</INSDQualifier_name>
      <INSDQualifier_value>isoguanine</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>
```
Example 2: Non post-translationally modified amino acid

Example 3: D-amino acid

31. 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.

32. 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.

33. The following examples illustrate the presentation of “unknown” or “other” amino acids according to paragraphs 31 and 32 above:

Example 1: “unknown” amino acid

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 65 to 71). For presentation of sequence variants, i.e., deletions, insertions, or substitutions, see paragraphs 92 to 97.

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 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 26, respectively), wherein the exact number of residues in each region is disclosed, must be included in the sequence listing as a single sequence with a single 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 be included in the sequence listing as multiple, separate sequences. Each such separate sequence must contain one region of specifically enumerated residues with its own sequence identification number, wherein 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 5 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 version="1.0" encoding="UTF-8"?>.
```

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

```
<!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;

(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 72 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.

_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 &quot;V#_#&quot;, e.g. &quot;V1_0&quot;.</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 &quot;CCYY-MM-DD&quot;).</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:

```
<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_

45. 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>ApplicationIdentification</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>The ApplicationIdentification is composed of:</td>
<td></td>
<td></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>EarliestPriorityApplicationIdentificaiton</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>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>Element</td>
<td>Description</td>
<td>Mandatory/Optional</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------------</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 9).</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" 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>
```

*{...}* represents relevant information for each sequence that has not been included in this example.
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>14/999,999</ApplicationNumberText>
  <FilingDate>2015-01-05</FilingDate>
</ApplicationIdentification>
<ApplicationIdentification>
  <IPOfficeCode>IB</IPOfficeCode>
  <ApplicationNumberText>PCT/IB2014/099999</ApplicationNumberText>
  <FilingDate>2014-07-10</FilingDate>
</ApplicationIdentification>
<ApplicantFileReference>AB123</ApplicantFileReference>
<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>
```

*{...} 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 paragraph 8 b) 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 b), 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 paragraph 8 b) 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 太郎</InventorName>
<InventorNameLatin>Taro Tokkyo</InventorNameLatin>
>InventionTitle languageCode="JA"> efg タンパク質のためのマウス abcd-1 遺伝子</InventionTitle>
>InventionTitle languageCode="EN"> Mus musculus abcd-1 gene for efg protein</InventionTitle>
>InventionTitle languageCode="FR"> Gène abcd-1 de Mus musculus pour protéine efg</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 9) for each sequence is contained. For example:

   <SequenceData sequenceIDNumber="1"/>

52. The SequenceData element must contain a dependent element INSDSeq, 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>INSDSeq_length</td>
<td>Length of the sequence</td>
<td>Mandatory</td>
</tr>
<tr>
<td>INSDSeq_moltype</td>
<td>Molecule type</td>
<td>Mandatory</td>
</tr>
<tr>
<td>INSDSeq_division</td>
<td>Indication that a sequence is related to a patent application</td>
<td>Mandatory with the value &quot;PAT&quot;</td>
</tr>
<tr>
<td>INSDSeq_feature-table</td>
<td>List of annotations of the sequence</td>
<td>Mandatory</td>
</tr>
<tr>
<td>INSDSeq_sequence</td>
<td>Sequence</td>
<td>Mandatory</td>
</tr>
</tbody>
</table>

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

   <INSDSeq_length>8</INSDSeq_length>

54. The element INSDSeq_moltype must disclose the type of molecule that is being presented. For nucleotide sequences, the molecule type must be indicated as DNA or RNA. For protein or polypeptide 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 85). For example:

   <INSDSeq_moltype>AA</INSDSeq_moltype>

55. Where a nucleotide sequence contains both DNA and RNA fragments, the value for INSDSeq_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:

```xml
<INSDSeq>
  <INSDSeq_length>120</INSDSeq_length>
  <INSDSeq_moltype>DNA</INSDSeq_moltype>
  <INSDSeq_division>PAT</INSDSeq_division>
  <INSDSeq_feature-table>
    <INSDFeature>
      <INSDFeature_key>source</INSDFeature_key>
      <INSDFeature_location>1..120</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>other DNA</INSDQualifier_value>
        </INSDQualifier>
      </INSDFeature_quals>
    </INSDFeature>
    <INSDFeature>
      <INSDFeature_key>misc_feature</INSDFeature_key>
      <INSDFeature_location>1..60</INSDFeature_location>
      <INSDFeature_quals>
        <INSDQualifier>
          <INSDQualifier_name>note</INSDQualifier_name>
          <INSDQualifier_value>DNA fragment</INSDQualifier_value>
        </INSDQualifier>
      </INSDFeature_quals>
    </INSDFeature>
    <INSDFeature>
      <INSDFeature_key>misc_feature</INSDFeature_key>
      <INSDFeature_location>61..120</INSDFeature_location>
      <INSDFeature_quals>
        <INSDQualifier>
          <INSDQualifier_name>note</INSDQualifier_name>
          <INSDQualifier_value>RNA fragment</INSDQualifier_value>
        </INSDQualifier>
      </INSDFeature_quals>
    </INSDFeature>
  </INSDSeq_feature-table>
  <INSDSeq_sequence>
    cgacccacgcgtccgaggaaccaaccatcacgtttgaggacttcgtgaaggaattggataatacccgctccctaccaaaatggcg
    agcgccgactcattgctcctcgtaccgtcgagcgc
  </INSDSeq_sequence>
</INSDSeq>
```

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). The sequence must not contain numbers, punctuation or whitespace characters.

58. An intentionally skipped sequence must be presented 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 presentation of an intentionally skipped sequence as per paragraph 58 above:

```xml
<SequenceData sequenceIDNumber="3">
  <INSSeq>
    <INSSeq_length/>
    <INSSeq_moltype/>
    <INSSeq_division/>
    <INSSeq_sequence>000</INSSeq_sequence>
  </INSSeq>
</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 `INSSeq_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 presented 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**

65. 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 `INSSeq_sequence` element, and may contain one or more location operator(s) (see paragraphs 68 to 71).

66. 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 68 to 71). The location descriptor must not include numbering for residues beyond the range of the sequence in the `INSSeq_sequence` element.

67. 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 `INSSeq_sequence` element, and x is less than y.
68. 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.</td>
</tr>
</tbody>
</table>

69. 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, may 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).

70. 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.

71. The following examples illustrate feature locations, as per paragraphs 65 to 70 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>Location Example</td>
<td>Description</td>
</tr>
<tr>
<td>------------------</td>
<td>-------------</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>
<tr>
<td>join(12..78,134..202)</td>
<td>Indicates that regions 12 to 78 and 134 to 202 should be joined to form one contiguous sequence.</td>
</tr>
</tbody>
</table>

(b) locations for nucleotides only:

<table>
<thead>
<tr>
<th>Location example</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>complement(34..126)</td>
<td>Start at the base complementary to 126 and finish at the base complementary to base 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 bases 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>

72. In an XML instance of a sequence listing, the characters "<" and ">") in a location descriptor must be replaced by the appropriate predefined entities (see paragraph 41). For example:

- Feature location "<1":
  `<INSDFeature_location>&lt;1</INSDFeature_location>`
- Feature location "1..>888":
  `<INSDFeature_location>1..&gt;888</INSDFeature_location>`

**Feature qualifiers**

73. 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 86 and 87);
(b) controlled vocabulary or enumerated values (e.g. a number or date); and
(c) sequences.

74. 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.

75. Any sequence encompassed by paragraph 6 which is provided as a qualifier value must be separately listed in the sequence listing with its own sequence identification number.

**Mandatory feature qualifiers**

76. 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**

77. 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:
78. 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.

79. 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.

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

Example 1: Source for a nucleotide sequence

```xml
<INSDSeq_feature-table>
    <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..5164</INSDFeature_location>
        <INSDFeature_quals>
            <INSDQualifier>
                <INSDQualifier_name>organism</INSDQualifier_name>
                <INSDQualifier_value>Solanum lycopersicum</INSDQualifier_value>
            </INSDQualifier>
            <INSDQualifier>
                <INSDQualifier_name>note</INSDQualifier_name>
                <INSDQualifier_value>common name: tomato</INSDQualifier_value>
            </INSDQualifier>
            <INSDQualifier>
                <INSDQualifier_name>mol_type</INSDQualifier_name>
                <INSDQualifier_value>genomic DNA</INSDQualifier_value>
            </INSDQualifier>
        </INSDFeature_quals>
    </INSDFeature>
</INSDSeq_feature-table>
```

Example 2: Source for a protein sequence

```xml
<INSDSeq_feature-table>
    <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..174</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>
```

81. 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:

```xml
<INSDQualifier_name>organism</INSDQualifier_name>
<INSDQualifier_value>Bacillus sp.</INSDQualifier_value>
```
82. If the source of the sequence is natural, 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:

```
<INSDQualifier_name>organism</INSDQualifier_name>
<INSDQualifier_value>unidentified</INSDQualifier_value>
<INSDQualifier_name>note</INSDQualifier_name>
<INSDQualifier_value>bacterium B8</INSDQualifier_value>
```

83. 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:

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

84. 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:

```
<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>
```

85. 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 INSSeq_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”, “tRNA”, “rRNA”, “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

86. 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.

87. 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

88. 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 INSFeature_location should identify the 
location of the “CDS” feature and must include the stop codon.

89. 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.
90. A protein sequence encoded by the coding sequence and disclosed in a “translation” qualifier that is encompassed by paragraph 6 must be assigned its own sequence identification number and be presented in the sequence listing. The sequence identification number assigned to the protein 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 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>MLVHLERTTIMDFSSLINLILWGLLIIAIAYILMLGDFDMILGMLPAPSAKCRDHMISSISAPFWDGNETWLVGGGGLFAA
            FPLAYSILMPAFYIPIIIMLGLIVFVSEFRFLKEGKRYRRLWDIFAFHFSLGAACQGMFLGAPFHGEVNGRNFSGGQLM</INSDQualifier_value>
      </INSDQualifier>
      <INSDQualifier>
        <INSDQualifier_name>protein_id</INSDQualifier_name>
        <INSDQualifier_value>89</INSDQualifier_value>
      </INSDQualifier>
    </INSDFeature_quals>
  </INSDFeature>
</INSDSeq_feature-table>
```

Variants

91. A primary sequence and any variant of that sequence, each disclosed by enumeration of their residues and encompassed by paragraph 6 must be presented in the sequence listing with their own sequence identification number.

92. 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 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 its own sequence identification number, where it contains variations at multiple distinct locations and the occurrence of those variations are interdependent; and

   (c) must be presented as a separate sequence with its own sequence identification number, where it contains an inserted or substituted sequence that contains in excess of 1000 residues (see paragraph 87).

93. 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</td>
<td>Naturally occurring mutations and polymorphisms, eg., Alleles, RFLPs.</td>
</tr>
<tr>
<td>Nucleic acid</td>
<td>misc_difference</td>
<td>replace</td>
<td>Variability introduced artificially, e.g., by genetic manipulation or by chemical synthesis.</td>
</tr>
<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>
94. 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. A deletion must be represented by an empty qualifier value. An inserted or substituted residue(s) must be provided in the “replace” or “NOTE” qualifier. The value format for the “replace” and “NOTE” qualifiers is free text and must not exceed 1000 characters, as provided in paragraph 87. See paragraph 97 for sequences encompassed by paragraph 6 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.

95. 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.

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

Example 1: Feature key “variation” for a substitution in a nucleotide sequence. A cytosine replaces the nucleotide given in position 413 of the sequence.

```
<INSDFeature>
  <INSDFeature_key>variation</INSDFeature_key>
  <INSDFeature_location>413</INSDFeature_location>
  <INSDFeature_quals>
    <INSDQualifier>
      <INSDQualifier_name>replace</INSDQualifier_name>
      <INSDQualifier_value>c</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>
```

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

```
<INSDFeature>
  <INSDFeature_key>misc_difference</INSDFeature_key>
  <INSDFeature_location>413</INSDFeature_location>
  <INSDFeature_quals>
    <INSDQualifier>
      <INSDQualifier_name>replace</INSDQualifier_name>
      <INSDQualifier_value></INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>
```

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.

```
<INSDFeature>
  <INSDFeature_key>misc_difference</INSDFeature_key>
  <INSDFeature_location>100^101</INSDFeature_location>
  <INSDFeature_quals>
    <INSDQualifier>
      <INSDQualifier_name>replace</INSDQualifier_name>
      <INSDQualifier_value>atgccaaatat</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>
```

Example 4: 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 I, A, F, Y, alIle, MeIle, or Nle.

```
<INSDFeature>
  <INSDFeature_key>VARIANT</INSDFeature_key>
  <INSDFeature_location>100</INSDFeature_location>
  <INSDFeature_quals>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>I, A, F, Y, aIle, MeIle, or Nle</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>
```
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.

<INSDFeature>
  <INSDFeature_key>VARIANT</INSDFeature_key>
  <INSDFeature_location>100</INSDFeature_location>
  <INSDFeature_quals>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>not K, R, or H</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>

97. A sequence encompassed by paragraph 6 that is provided as an insertion or a substitution in a qualifier value for a primary sequence annotation must also be presented in the sequence listing with its own sequence identification number.
ST.26 - ANNEX I

CONTROLLED VOCABULARY

Final Draft

Proposal presented by the SEQL Task Force for consideration and adoption at the CWS/4

TABLE OF CONTENTS

SECTION 1: LIST OF NUCLEOTIDES ................................................................. 23
SECTION 2: LIST OF MODIFIED NUCLEOTIDES ........................................ 23
SECTION 3: LIST OF AMINO ACIDS ............................................................. 25
SECTION 4: LIST OF MODIFIED AND UNUSUAL AMINO ACIDS ............... 26
SECTION 5: FEATURES KEYS FOR NUCLEIC SEQUENCES ...................... 27
SECTION 6: DESCRIPTION OF QUALIFIERS FOR NUCLEIC SEQUENCES ................................................................. 47
SECTION 7: FEATURE KEYS FOR AMINO ACID SEQUENCES .................. 65
SECTION 8: QUALIFIERS FOR AMINO ACID SEQUENCES ......................... 71
SECTION 9: GENETIC CODES TABLES .......................................................... 72
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>cmmn5s2u</td>
<td>5-carboxymethylaminomethyl-2-thiouridine</td>
</tr>
<tr>
<td>cmmn5u</td>
<td>5-carboxymethylaminomethyluridine</td>
</tr>
<tr>
<td>d</td>
<td>dihydrouridine</td>
</tr>
<tr>
<td>fm</td>
<td>2’-O-methylpseudouridine</td>
</tr>
<tr>
<td>gal q</td>
<td>beta-D-galactosylqueosine</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>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>Abbreviation</td>
<td>Modified Nucleotide</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>mam5s2u</td>
<td>5-methoxyaminomethyl-2-thiouridine</td>
</tr>
<tr>
<td>man q</td>
<td>beta,D-mannosylqueosine</td>
</tr>
<tr>
<td>mcm5s2u</td>
<td>5-methoxycarbonylmethyl-2-thiouridine</td>
</tr>
<tr>
<td>mcm5u</td>
<td>5-methoxycarbonylmethyluridine</td>
</tr>
<tr>
<td>mo5u</td>
<td>5-methoxyuridine</td>
</tr>
<tr>
<td>ms26a</td>
<td>2-methylthio-N6-isopentenyladenosine</td>
</tr>
<tr>
<td>ms2t6a</td>
<td>N-((9-beta-D-ribofuranosyl-2-methylthiopurine-6-yl)carbamoyl)threonine</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-oxyacetic 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>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 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 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", 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>Pyrrolysine</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</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-Aminoadipic acid</td>
</tr>
<tr>
<td>bAad</td>
<td>3-Aminoadipic 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>alle</td>
<td>allo-Isoleucine</td>
</tr>
<tr>
<td>MeGly</td>
<td>N-Methylglycine, sarcosine</td>
</tr>
<tr>
<td>MeLle</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>
SECTION 5: FEATURES KEYS FOR NUCLEIC SEQUENCES

This paragraph contains the list of allowed feature keys to be used for nucleotide 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; when a parent key is indicated in the description of a feature key, it is 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.

5.1. Feature Key attenuator

Definition 1) region of DNA at which regulation of termination of transcription occurs, which controls the expression of some bacterial operons; 2) sequence segment located between the promoter and the first structural gene that causes partial termination of transcription

Optional qualifiers allele
gene
gene_synonym
map
note
operon
phenotype

Organism scope prokaryotes

Molecule scope DNA

5.2. Feature Key C_region

Definition 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

Optional qualifiers allele
gene
gene_synonym
map
note
product
pseudo
pseudogene
standard_name

Parent Key CDS

Organism scope eukaryotes
5.3. Feature Key CAAT_signal

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

Optional qualifiers:
allele
gene
gene_synonym
map
note

Organism scope:
eukaryotes and eukaryotic viruses

Molecule scope:
DNA

References:

5.4. 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
number
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 and 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 amino acids
### 5.5. 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.

### 5.6. 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.

### 5.7. 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.

**Parent Key**: CDS.

**Organism scope**: Eukaryotes.

### 5.8. Feature Key: enhancer

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

**Optional qualifiers**: allele, bound moiety, gene, gene synonym, map, note, standard name.

**Organism scope**: Eukaryotes and eukaryotic viruses.
5.9. 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, standard_name, trans_splicing

---

5.10. Feature Key: GC_signal

**Definition:** GC box; a conserved GC-rich region located upstream of the start point of eukaryotic transcription units which may occur in multiple copies or in either orientation; consensus=GGCCGG

**Optional qualifiers:** allele, gene, gene_synonym, map, note

**Organism scope:** Eukaryotes and eukaryotic viruses

---

5.11. 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.
<table>
<thead>
<tr>
<th>Feature Key</th>
<th>Definition</th>
<th>Optional qualifiers</th>
<th>Molecule scope</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>5.12. Feature Key</strong></td>
<td><strong>iDNA</strong></td>
<td>intervening DNA; DNA which is eliminated through any of several kinds of recombination</td>
<td>allele, function, gene, gene_synonym, map, note, number, standard_name</td>
<td>DNA</td>
</tr>
<tr>
<td><strong>5.13. Feature Key</strong></td>
<td><strong>intron</strong></td>
<td>a segment of DNA that is transcribed, but removed from within the transcript by splicing together the sequences (exons) on either side of it</td>
<td>allele, function, gene, gene_synonym, map, note, number, pseudo, pseudogene, standard_name, trans_splicing</td>
<td></td>
</tr>
<tr>
<td><strong>5.14. Feature Key</strong></td>
<td><strong>J_segment</strong></td>
<td>joining segment of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains</td>
<td>allele, gene, gene_synonym, map, note, product, pseudo, pseudogene, standard_name</td>
<td></td>
</tr>
<tr>
<td><strong>5.15. Feature Key</strong></td>
<td><strong>LTR</strong></td>
<td>long terminal repeat, a sequence directly repeated at both ends of a defined sequence, of the sort typically found in retroviruses</td>
<td>allele, function, gene, gene_synonym, map, note, standard_name</td>
<td></td>
</tr>
</tbody>
</table>
5.16. 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

5.17. 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 feature key RBS is used for ribosome binding sites

5.18. 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 should 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.
5.19. 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

5.20. 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 standard_name
Molecule scope DNA

5.21. 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 \textit{misc\_signal}

\textbf{Definition}: any region containing a signal controlling or altering gene function or expression that cannot be described by other signal keys (promoter, CAAT\_signal, TATA\_signal, -35\_signal, -10\_signal, GC\_signal, RBS, polyA\_signal, enhancer, attenuator, terminator, and rep\_origin).

\textbf{Optional qualifiers}: allele, function, gene, gene\_synonym, map, note, operon, phenotype, standard\_name.

5.23. Feature Key \textit{misc\_structure}

\textbf{Definition}: any secondary or tertiary nucleotide structure or conformation that cannot be described by other Structure keys (stem\_loop and D-loop).

\textbf{Optional qualifiers}: allele, function, gene, gene\_synonym, map, note, standard\_name.

5.24. Feature Key \textit{mobile\_element}

\textbf{Definition}: region of genome containing mobile elements.

\textbf{Mandatory qualifiers}: mobile\_element\_type.

\textbf{Optional qualifiers}: allele, function, gene, gene\_synonym, map, note, rpt\_family, rpt\_type, standard\_name.

5.25. Feature Key \textit{modified\_base}

\textbf{Definition}: the indicated nucleotide is a modified nucleotide and should be substituted for by the indicated molecule (given in the mod\_base qualifier value).

\textbf{Mandatory qualifiers}: mod\_base.

\textbf{Optional qualifiers}: allele, frequency, gene, gene\_synonym, map, note.

\textbf{Comment}: value for the mandatory mod\_base qualifier is limited to the restricted vocabulary for modified base abbreviations in Section 2 of this Annex.
### 5.26. 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
- artificial_location
- function
- gene
- gene_synonym
- map
- note
- operon
- product
- pseudo
- pseudogene
- standard_name
- trans_splicing

### 5.27. 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
- function
- gene
- gene_synonym
- map
- note
- operon
- product
- pseudo
- pseudogene
- standard_name
- trans_splicing

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

### 5.28. Feature Key N_region

**Definition**
extra nucleotides inserted between rearranged immunoglobulin segments

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

**Parent Key**
CDS

**Organism scope**
eukaryotes
5.29. Feature Key operon
Definition region containing polycistronic transcript including a cluster of genes that are under the control of the same regulatory sequences/promotor and in the same biological pathway
Mandatory qualifiers operon
Optional qualifiers allele function map note phenotype pseudo pseudogene standard_name

5.30. 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 should be used for origins of replication; direction qualifier has legal values RIGHT, LEFT and BOTH, however only RIGHT and LEFT 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
5.32. Feature Key  
**polyA_site**

**Definition**
A 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

5.33. Feature Key  
**precursor_RNA**

**Definition**
Any RNA species that is not yet the mature RNA product; may include 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.

5.34. Feature Key  
**prim_transcript**

**Definition**
Primary (initial, unprocessed) transcript; includes 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
- standard_name

5.35. Feature Key  
**primer_bind**

**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 times may be stored under the PCR_conditions qualifier; since PCR reactions most often involve pairs of primers, a single primer_bind key may use the order(location,location) operator with two locations, or a pair of primer_bind keys may be used.
5.36. Feature Key promoter
Definition region on a DNA molecule involved in RNA polymerase binding to initiate transcription
Optional qualifiers allele bound_moiety function gene gene_synonym map note operon phenotype pseudo pseudogene standard_name
Molecule scope DNA

5.37. Feature Key protein_bind
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 RBS is used for ribosome binding sites

5.38. Feature Key RBS
Definition ribosome binding site
Optional qualifiers allele gene gene_synonym map note standard_name
Comment in prokaryotes, known as the Shine-Dalgarno sequence: is located 5 to 9 bases upstream of the initiation codon; consensus GGAGGT [1,2]
5.39. Feature Key repeat_region
Definition region of genome containing repeating units
Optional qualifiers allele
function
gene
gene_synonym
map
note
rpt_family
rpt_type
rpt_unit_range
rpt_unit_seq
satellite
standard_name

5.40. Feature Key rep_origin
Definition origin of replication; starting site for duplication of nucleic acid to give two identical copies
Optional Qualifiers allele
direction
gene
gene_synonym
map
note
standard_name
Comment direction qualifier has valid values: RIGHT, LEFT, or BOTH

5.41. 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
standard_name
Comment rRNA sizes should be annotated with the product qualifier
<table>
<thead>
<tr>
<th>Feature Key</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>s_region</td>
<td>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</td>
</tr>
<tr>
<td>sig_peptide</td>
<td>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</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Optional qualifiers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>allele</td>
<td></td>
</tr>
<tr>
<td>gene</td>
<td></td>
</tr>
<tr>
<td>gene_synonym</td>
<td></td>
</tr>
<tr>
<td>map</td>
<td></td>
</tr>
<tr>
<td>note</td>
<td></td>
</tr>
<tr>
<td>product</td>
<td></td>
</tr>
<tr>
<td>pseudo</td>
<td></td>
</tr>
<tr>
<td>pseudogene</td>
<td></td>
</tr>
<tr>
<td>standard_name</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parent Key</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>misc_signal</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Organism scope</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>eukaryotes</td>
<td></td>
</tr>
</tbody>
</table>
5.44. 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

5.45. 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
5.46. Feature Key
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

5.47. Feature Key
Definition TATA box; Goldberg-Hogness box; a conserved AT-rich septamer found about 25 bp
before the start point of each eukaryotic RNA polymerase II transcript unit which
may be involved in positioning the enzyme for correct initiation; consensus=TATA(A
or T)A(A or T) [1,2]
Optional qualifiers allele
gene
gene_synonym
map
note
Organism scope eukaryotes and eukaryotic viruses
Molecule scope DNA
Science 209, 1406-1414 (1980)

5.48. Feature Key
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
gen_synonym
map
note
operon
standard_name
Molecule scope DNA

5.50. 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
gen_synonym
map
note
product
pseudo
pseudogene
standard_name
tag_peptide

5.51. Feature Key transit_peptide
Definition transit peptide coding sequence; coding sequence for an N-terminal domain of a nuclear-encoded organellar protein; this domain is involved in post-translational import of the protein into the organelle
Optional qualifiers allele
function
gene
gen_synonym
map
note
product
pseudo
pseudogene
standard_name

5.52. 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
gen_synonym
map
note
product
pseudo
pseudogene
standard_name
trans_splicing
5.53. Feature Key

Definition

author is unsure of exact sequence in this region

Optional qualifiers

allele
compare
gene
gene_synonym
map
note
replace

Comment

use the replace qualifier to annotate a deletion, insertion, or substitution.

5.54. Feature Key

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

5.55. Feature Key

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
note
product
pseudo
pseudogene
standard_name

Parent Key

CDS

Organism scope
eukaryotes
5.56. 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 note phenotype product replace standard_name
Comment used to describe alleles, RFLPs, and other naturally occurring mutations and polymorphisms; variability arising as a result of genetic manipulation (e.g. site directed mutagenesis) should be described with the misc_difference feature; use the replace qualifier to annotate a deletion, insertion, or substitution

5.57. Feature Key 3'UTR
Definition region at the 3' end of a mature transcript (following the stop 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 “3'UTR” must be represented as “3&amp;apos;UTR” in the XML file, i.e., <INSDFeature_key>3&apos;UTR</INSDFeature_key>.

5.58. Feature Key 5'UTR
Definition region at the 5' end of a mature transcript (preceding the 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&amp;apos;UTR” in the XML file, i.e., <INSDFeature_key>5&apos;UTR</INSDFeature_key>.
### 5.59. Feature Key: -10_signal

**Definition**
Prbnow box; a conserved region about 10 bp upstream of the start-point of bacterial transcription units which may be involved in binding RNA polymerase; consensus=TATAAT [1,2,3,4]

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

**Organism scope**
prokaryotes

**Molecule scope**
DNA

**References**

### 5.60. Feature Key: -35_signal

**Definition**
a conserved hexamer about 35 bp upstream of the start-point of bacterial transcription units; consensus=TTGACA or TGTTGACA

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

**Organism scope**
prokaryotes

**Molecule scope**
DNA

**References**
SECTION 6: DESCRIPTION OF QUALIFIERS FOR NUCLEIC SEQUENCES

This section contains the list of qualifiers to be used for features in nucleotide 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.

6.1. Qualifier allele
Definition name of the allele for the given gene
Value format free text
Example <INSDQualifier_value>adh1-1</INSDQualifier_value>
Comment 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.

6.2. Qualifier anticodon
Definition location of the anticodon of tRNA and the amino acid for which it codes
Value format (pos:<location>,aa:<amino_acid>,seq:<text>) where location is the position of the anticodon and <amino_acid> is the abbreviation for the amino acid encoded and seq is the sequence of the anticodon
Example <INSDQualifier_value>(pos:34..36,aa:Phe,seq:aaa)</INSDQualifier_value> <INSDQualifier_value>(pos:join(5,495..496,aa:Leu,seq:taa)</INSDQualifier_value> <INSDQualifier_value>(pos:complement(4156..4158),aa:Glu,seq:ttg)</INSDQualifier_value>

6.3. Qualifier bound_moiety
Definition name of the molecule/complex that may bind to the given feature
Value format free text
Example <INSDQualifier_value>GAL4</INSDQualifier_value>
Comment 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.

6.4. Qualifier cell_line
Definition cell line from which the sequence was obtained
Value format free text
Example <INSDQualifier_value>MCF7</INSDQualifier_value>

6.5. Qualifier cell_type
Definition cell type from which the sequence was obtained
Value format free text
Example <INSDQualifier_value>leukocyte</INSDQualifier_value>
6.6. Qualifier chromosome
Definition chromosome (e.g. Chromosome number) from which the sequence was obtained
Value format free text
Example <INSDQualifier_value>1</INSDQualifier_value> <INSDQualifier_value>X</INSDQualifier_value>

6.7. Qualifier clone
Definition clone from which the sequence was obtained
Value format free text
Example <INSDQualifier_value>lambda-hIL7.3</INSDQualifier_value>
Comment not more than one clone should be specified for a given source feature; 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.

6.8. Qualifier clone_lib
Definition clone library from which the sequence was obtained
Value format free text
Example <INSDQualifier_value>lambda-hIL7</INSDQualifier_value>

6.9. Qualifier codon_start
Definition indicates the offset at which the first complete codon of a coding feature can be found, relative to the first base of that feature.
Value format 1 or 2 or 3
Example <INSDQualifier_value>2</INSDQualifier_value>

6.10. Qualifier collected_by
Definition name of persons or institute who collected the specimen
Value format free text
Example <INSDQualifier_value>Dan Janzen</INSDQualifier_value>

6.11. Qualifier collection_date
Definition date that the specimen was collected
Value format DD-Mmm-YYYY, Mmm-YYYY or YYYY
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 abbreviation 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**
<Nipponbare>
<Tenuifolius>
<Candy Cane>
<IR36>

**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**
<fourth instar larva>

### 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**
<LEFT>

**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 can be replaced by dash "-" to indicate uncertain assignment. Symbol "n" can 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. Sequences including environmental_sample 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>rarrangement required for product</INSDQualifier_value>`
`<INSDQualifier_value>annotated by transcript or proteomic data</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 according 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

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

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.

6.22. Qualifier gene

Definition  symbol of the gene corresponding to a sequence region

Value format  free text

Example  

<INSDQualifier_value>ilvE</INSDQualifier_value>

Comment  Use gene qualifier to provide the gene symbol; use standard_name qualifier to provide the full gene name.

6.23. Qualifier gene_synonym

Definition  synonymous, replaced, obsolete or former gene symbol

Value format  free text

Example  

<INSDQualifier_value>Hox-3.3</INSDQualifier_value>

in a feature where the gene qualifier value is Hoxc6

Comment  used where it is helpful to indicate a gene symbol synonym; when used, a primary gene symbol must always be indicated in a gene qualifier

6.24. Qualifier germline

Definition  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

Value format  none

Comment  germline qualifier should not be used to indicate that the source of the sequence is a gamete or germ cell; germline and rearranged qualifiers cannot be used in the same source feature; germline and rearranged qualifiers should 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)
6.25. **Qualifier** haplogroup

**Definition** name for a group of similar haplotypes that share some sequence variation. Haplogroups are often used to track migration of population groups.

**Value format** free text

**Example** <INSDQualifier_value>H*</INSDQualifier_value>

6.26. **Qualifier** haplotype

**Definition** 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.

**Value format** free text

**Example** <INSDQualifier_value>Dw3 B5 Cw1 A1</INSDQualifier_value>

6.27. **Qualifier** host

**Definition** natural (as opposed to laboratory) host to the organism from which sequenced molecule was obtained

**Value format** free text

**Example** <INSDQualifier_value>Homo sapiens</INSDQualifier_value> <INSDQualifier_value>Homo sapiens 12 year old girl</INSDQualifier_value> <INSDQualifier_value>Rhizobium NGR234</INSDQualifier_value>

6.28. **Qualifier** identified_by

**Definition** name of the expert who identified the specimen taxonomically

**Value format** free text

**Example** <INSDQualifier_value>John Burns</INSDQualifier_value>

6.29. **Qualifier** isolate

**Definition** individual isolate from which the sequence was obtained

**Value format** free text

**Example** <INSDQualifier_value>Patient #152</INSDQualifier_value> <INSDQualifier_value>DGGE band PSBAC-13</INSDQualifier_value>

6.30. **Qualifier** isolation_source

**Definition** describes the physical, environmental and/or local geographical source of the biological sample from which the sequence was derived

**Value format** free text

**Examples** <INSDQualifier_value>rumen isolates from standard Pelleted ration-fed steer #67</INSDQualifier_value> <INSDQualifier_value>permanent Antarctic sea ice</INSDQualifier_value> <INSDQualifier_value>denitrifying activated sludge from carbon_limited continuous reactor</INSDQualifier_value>

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

**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[d.dddd] N|S d[d.dddd] W|E".

**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.

**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.

**Examples**
- `<INSDQualifier_value>MAT-1</INSDQualifier_value>`
- `<INSDQualifier_value>plus</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** `<mobile_element_type>::<mobile_element_name>`  
where `<mobile_element_type>` is one of the following:  
transposon  
retrotransposon  
integron  
isolation sequence  
non-LTR retrotransposon  
SI NE  
MITE  
LINE  
other  
**Example** `<INSDQualifier_value>transposon:Tnp9</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, Table 2  
**Example** `<INSDQualifier_value>m5c</INSDQualifier_value>`  
**Other** `<INSDQualifier_value>OTHER</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.

6.38. **Qualifier** mol_type  
**Definition** molecule type of sequence  
**Value format** One chosen from the following:  
genomic DNA  
genomic RNA  
rRNA  
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 should be described using "genomic DNA"); ribosomal RNA genes should be described using "genomic DNA"; "rRNA" should only be used if the ribosomal RNA molecule itself has been sequenced; values "other RNA" and "other DNA" should be applied to synthetic molecules, values "unassigned DNA", "unassigned RNA" should be applied where in vivo molecule is unknown.
### 6.39. Qualifier

**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`
- `raslRNA`
- `scRNA`
- `siRNA`
- `miRNA`
- `piRNA`
- `snoRNA`
- `snRNA`
- `SRP_RNA`
- `vault_RNA`
- `Y_RNA`
- `other`

**Example**

```xml
<INSDQualifier_value>autocatalytically_spliced_intron</INSDQualifier_value>
<INSDQualifier_value>siRNA</INSDQualifier_value>
<INSDQualifier_value>scRNA</INSDQualifier_value>
<INSDQualifier_value>other</INSDQualifier_value>
```

**Comment**

Specific ncRNA types not yet in the ncRNA_class controlled vocabulary can be annotated by entering "other" as the ncRNA_class qualifier value, and providing a brief explanation of novel ncRNA_class in a note qualifier.

### 6.40. Qualifier

**note**

**Definition**

Any comment or additional information.

**Value format**

Free text.

**Example**

```xml
<INSDQualifier_value>A comment about the feature</INSDQualifier_value>
```

### 6.41. 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).

**Example**

```xml
<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.
### 6.42. Qualifier

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

**Value format**
free text

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

**Comment**
valid only on Prokaryota-specific features

### 6.43. Qualifier

**Definition**
type of membrane-bound intracellular structure from which the sequence was obtained

**Value format**
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

**Examples**
- <INSDQualifier_value>chromatophore</INSDQualifier_value>
- <INSDQualifier_value>hydrogenosome</INSDQualifier_value>
- <INSDQualifier_value>mitochondrion</INSDQualifier_value>
- <INSDQualifier_value>nucleomorph</INSDQualifier_value>
- <INSDQualifier_value>plastid</INSDQualifier_value>
- <INSDQualifier_value>mitochondrion:kinetoplast</INSDQualifier_value>
- <INSDQualifier_value>plastid:chloroplast</INSDQualifier_value>
- <INSDQualifier_value>plastid:apicoplast</INSDQualifier_value>
- <INSDQualifier_value>plastid:chromoplast</INSDQualifier_value>
- <INSDQualifier_value>plastid:cyanelle</INSDQualifier_value>
- <INSDQualifier_value>plastid:leucoplast</INSDQualifier_value>
- <INSDQualifier_value>plastid:proplastid</INSDQualifier_value>

### 6.44. Qualifier

**Definition**
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

**Value format**
free text

**Example**
<INSDQualifier_value>Homo sapiens</INSDQualifier_value>
6.45. Qualifier  
PCR_primers

**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**
[fwd_name: XXX1, fwd_seq: xxxxx1, fwd_name: XXX2, fwd_seq: xxxxx2, rev_name: YYYY1, rev_seq: yyyyy1, rev_name: YYYY2, rev_seq: yyyyy2]

**Example**

<INSDQualifier_value>fwd_name: CO1P1, fwd_seq: ttgattttttggtcayccwgaagt, rev_name: CO1R4, rev_seq: ccwvytardcctarraartgttg</INSDQualifier_value>

**Comment**
Both sequences should be presented in 5'→3' order. The sequences should be given in the symbols from Section 1 of this Annex, except for the modified bases; those 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  
Phenotype

**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

**Example**

<INSDQualifier_value>erythromycin resistance</INSDQualifier_value>

6.47. Qualifier  
Plasmid

**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

**Example**

<INSDQualifier_value>pC589</INSDQualifier_value>

6.48. Qualifier  
Pop_variant

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

**Value format**
free text

**Example**

<INSDQualifier_value>pop1</INSDQualifier_value>
<INSDQualifier_value>Bear Paw</INSDQualifier_value>

6.49. 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)
### 6.50. 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.

**Value format**
an integer greater than zero

**Example**

```xml
<INSDQualifier_value>89</INSDQualifier_value>
```

### 6.51. 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**
one

### 6.52. Qualifier

**pseudo**

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

**Value format**
one

**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 pseudogenisation 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

**pseudogene**

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

**Value format**

<table>
<thead>
<tr>
<th>TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>processed</td>
</tr>
<tr>
<td>unprocessed</td>
</tr>
<tr>
<td>unitary</td>
</tr>
<tr>
<td>allelic</td>
</tr>
<tr>
<td>unknown</td>
</tr>
</tbody>
</table>

**Example**

```xml
<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 mutation. 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 rearranged

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

Value format none

Comment The rearranged qualifier should not be used to annotate chromosome rearrangements that are not involved in an adaptive immune response; germline and rearranged qualifiers cannot be used in the same source feature; germline and rearranged qualifiers should 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 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 &lt;INSDQualifier_value&gt;a&lt;/INSDQualifier_value&gt; &lt;INSDQualifier_value&gt;&lt;/INSDQualifier_value&gt; - for a deletion

6.56. 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)] should be used in the CDS spans to indicate the location of ribosomal_slippage

6.57. Qualifier rpt_family

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

Value format free text

Example &lt;INSDQualifier_value&gt;Alu&lt;/INSDQualifier_value&gt;
6.58. Qualifier rpt_type

Definition organization of repeated sequence

Value format One of the following controlled vocabulary terms:
- tandem
- inverted
- flanking
- terminal
- direct
- dispersed
- other

Example <INSDQualifier_value>INVERTED</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;
- inverted - a repeat which occurs as part of as set (normally a part) organized in the reverse orientation;
- flanking - a repeat lying outside the sequence for which it has functional significance (eg. transposon insertion target sites);
- terminal - a repeat at the ends of and within the sequence for which it has functional significance (eg. transposon LTRs);
- direct - a repeat that exists not always adjacent but is in the same orientation;
- dispersed - a repeat that is found dispersed throughout the genome;
- other - a repeat exhibiting important attributes that cannot be described by other values.

6.59. Qualifier rpt_unit_range

Definition location (range) of a repeating unit

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.

6.60. Qualifier rpt_unit_seq

Definition identity of a repeat sequence

Value format free text

Example <INSDQualifier_value>aagggc</INSDQualifier_value>
<INSDQualifier_value>ag(5)tg(8)</INSDQualifier_value>
<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.
### 6.61. Qualifier: Satellite

**Definition**: Identifier for a satellite DNA marker, composed of many tandem repeats (identical or related) of a short basic repeated unit.

**Value format**: `<satellite_type>[:<class>][<identifier>]` - where `<satellite_type>` is one of the following:
- satellite
- microsatellite
- minisatellite

**Example**: `<INSDQualifier_value>satellite: S1a</INSDQualifier_value>`, `<INSDQualifier_value>satellite: alpha</INSDQualifier_value>`, `<INSDQualifier_value>satellite: gamma III</INSDQualifier_value>`, `<INSDQualifier_value>microsatellite: DC130</INSDQualifier_value>`

**Comment**: Many satellites have base composition or other properties that differ from those of the rest of the genome that allows them to be identified.

### 6.62. Qualifier: Segment

**Definition**: Name of viral or phage segment sequenced.

**Value format**: free text

**Example**: `<INSDQualifier_value>6</INSDQualifier_value>`

### 6.63. Qualifier: Serotype

**Definition**: Serological variety of a species characterized by its antigenic properties.

**Value format**: free text

**Example**: `<INSDQualifier_value>B1</INSDQualifier_value>`

**Comment**: 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 "Infraspecific Terms".

### 6.64. Qualifier: Serovar

**Definition**: Serological variety of a species (usually a prokaryote) characterized by its antigenic properties.

**Value format**: free text

**Example**: `<INSDQualifier_value>O157:H7</INSDQualifier_value>`

**Comment**: 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 "Infraspecific Terms".
6.65. **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**
- `<INSDQualifier_value>female</INSDQualifier_value>`
- `<INSDQualifier_value>male</INSDQualifier_value>`
- `<INSDQualifier_value>hermaphrodite</INSDQualifier_value>`
- `<INSDQualifier_value>unisexual</INSDQualifier_value>`
- `<INSDQualifier_value>bisexual</INSDQualifier_value>`
- `<INSDQualifier_value>asexual</INSDQualifier_value>`
- `<INSDQualifier_value>monoecious</INSDQualifier_value>` [or monecious]
- `<INSDQualifier_value>dioecious</INSDQualifier_value>` [or diecious]

**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.

6.66. **Qualifier** standard_name

**Definition**
accepted standard name for this feature

**Value format** free text

**Example**
`<INSDQualifier_value>dotted</INSDQualifier_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).

6.67. **Qualifier** strain

**Definition**
strain from which sequence was obtained

**Value format** free text

**Example**
`<INSDQualifier_value>BALB/c</INSDQualifier_value>`

**Comment**
entries including strain qualifier must not include the environmental_sample qualifier

6.68. **Qualifier** sub_clone

**Definition**
sub-clone from which sequence was obtained

**Value format** free text

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

**Comment**
not more than one sub_clone should be specified for a given source feature; to indicate that the sequence was obtained from multiple sub_clones, multiple source features should be given

6.69. **Qualifier** sub_species

**Definition**
name of sub-species of organism from which sequence was obtained

**Value format** free text

**Example**
`<INSDQualifier_value>lactis</INSDQualifier_value>`
6.70. Qualifier | sub_strain
---|---
**Definition** | 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
**Value format** | free text
**Example** | `<INSDQualifier_value>abis</INSDQualifier_value>`
**Comment** | If the parental strain is not given, this should be annotated in the strain qualifier instead of sub_strain. For example, either a strain qualifier with the value K-12 and a substrain qualifier with the value MG1655 or a strain qualifier with the value MG1655

6.71. Qualifier | tag_peptide
---|---
**Definition** | base location encoding the polypeptide for proteolysis tag of tmRNA and its termination codon
**Value format** | `<base_range>` - where `<base_range>` provides the first and last base (separated by two dots) of the location for the proteolysis tag
**Example** | `<INSDQualifier_value>90..122</INSDQualifier_value>`
**Comment** | 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 <90..122

6.72. Qualifier | tissue_lib
---|---
**Definition** | tissue library from which sequence was obtained
**Value format** | free text
**Example** | `<INSDQualifier_value>tissue library 772</INSDQualifier_value>`

6.73. Qualifier | tissue_type
---|---
**Definition** | tissue type from which the sequence was obtained
**Value format** | free text
**Example** | `<INSDQualifier_value>liver</INSDQualifier_value>`
6.74. **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 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:1017,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 code 'Sec' (one letter code 'U' 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 mRNA'.

6.75. **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.

6.76. **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 should be used only when the splice event is indicated in the "join" operator, e.g. join(complement(69611..69724),139856..140087)

6.77. **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>MATFPPWYRGCASTPSLKGLIMCTW</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 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.
6.78. 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 variatas 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 should be annotated separately with the ACT_SITE feature key. The corresponding amino acid residue number should 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. 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. The type of linkage (C-, N- or O-linked) to the protein is indicated in the “NOTE” qualifier. Examples of values used in the “NOTE” qualifier: O-linked (GlcNAc); C-linked (Man); N-linked (GlcNAc...); and O-linked (Glc...).
7.5. Feature Key CHAIN
Definition Extent of a polypeptide chain in the mature protein
Optional qualifiers NOTE

7.6. Feature Key COILED
Definition Extent of a coiled-coil region
Optional qualifiers NOTE

7.7. Feature Key COMPBIAS
Definition Extent of a compositionally biased region
Optional qualifiers NOTE

7.8. Feature Key CONFLICT
Definition Different sources report differing sequences.
Optional qualifiers NOTE

7.9. Feature Key CROSSLNK
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-Cter in ubiquitin)"

7.10. 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 cysteine 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)"
<table>
<thead>
<tr>
<th>Feature Key</th>
<th>Definition</th>
<th>Mandatory qualifiers</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA_BIND</td>
<td>Extent of a DNA-binding region</td>
<td>NOTE</td>
<td>The nature of the DNA-binding region is given in the NOTE qualifier. Examples of values for the “NOTE” qualifier: “Homeobox” and “Myb 2”</td>
</tr>
<tr>
<td>DOMAIN</td>
<td>Extent of a domain, which is defined as a specific combination of secondary structures organized into a characteristic three-dimensional structure or fold</td>
<td>NOTE</td>
<td>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”</td>
</tr>
<tr>
<td>HELIX</td>
<td>Secondary structure: Helices, for example, Alpha-helix; 3(10) helix; or Pi-helix</td>
<td>NOTE</td>
<td>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.</td>
</tr>
<tr>
<td>INIT_MET</td>
<td>Initiator methionine</td>
<td>NOTE</td>
<td>The location descriptor in the feature location element is &quot;1&quot;. This feature key indicates the N-terminal methionine is cleaved off. This feature is not used when the initiator methionine is not cleaved off.</td>
</tr>
<tr>
<td>INTRAMEM</td>
<td>Extent of a region located in a membrane without crossing it</td>
<td>NOTE</td>
<td></td>
</tr>
<tr>
<td>LIPID</td>
<td>Covalent binding of a lipid moiety</td>
<td>NOTE</td>
<td>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-diaclylglycerol cysteine.”</td>
</tr>
<tr>
<td>Feature Key</td>
<td>Definition</td>
<td>Mandatory qualifiers</td>
<td>Optional qualifiers</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------</td>
<td>----------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>METAL</td>
<td>Binding site for a metal ion.</td>
<td>NOTE</td>
<td></td>
</tr>
<tr>
<td>MOD_RES</td>
<td>Posttranslational modification of a residue</td>
<td>NOTE</td>
<td></td>
</tr>
<tr>
<td>MOTIF</td>
<td>Short (up to 20 amino acids) sequence motif of biological interest</td>
<td>NOTE</td>
<td>NOTE</td>
</tr>
<tr>
<td>MUTAGEN</td>
<td>Site which has been experimentally altered by mutagenesis</td>
<td>NOTE</td>
<td></td>
</tr>
<tr>
<td>NON_STD</td>
<td>Non-standard amino acid</td>
<td>NOTE</td>
<td>NOTE</td>
</tr>
<tr>
<td>NON_TER</td>
<td>The residue at an extremity of the sequence is not the terminal residue</td>
<td>NOTE</td>
<td></td>
</tr>
<tr>
<td>NP_BIND</td>
<td>Extent of a nucleotide phosphate-binding region</td>
<td>NOTE</td>
<td></td>
</tr>
</tbody>
</table>
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, Table 4, or 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
<table>
<thead>
<tr>
<th>Feature Key</th>
<th>Definition</th>
<th>Optional qualifiers</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>STRAND</td>
<td>Secondary structure: Beta-strand; for example Hydrogen bonded beta-strand or residue in an isolated beta-bridge</td>
<td>NOTE</td>
<td>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.</td>
</tr>
<tr>
<td>TOPO_DOM</td>
<td>Topological domain</td>
<td>NOTE</td>
<td></td>
</tr>
<tr>
<td>TRANSMEM</td>
<td>Extent of a transmembrane region</td>
<td>NOTE</td>
<td></td>
</tr>
<tr>
<td>TRANSIT</td>
<td>Extent of a transit peptide (mitochondrion, chloroplast, thylakoid, cyanelle, peroxisome etc.)</td>
<td>NOTE</td>
<td></td>
</tr>
<tr>
<td>TURN</td>
<td>Secondary structure Turns, for example, H-bonded turn (3-turn, 4-turn or 5-turn)</td>
<td>NOTE</td>
<td>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.</td>
</tr>
<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>
</tbody>
</table>
### 7.38. Feature Key VAR_SEQ

**Definition**
Description of sequence variants produced by alternative splicing, alternative promoter usage, alternative initiation and ribosomal frameshifting

**Optional qualifiers**
NOTE

### 7.39. Feature Key ZN_FING

**Definition**
Extent of a zinc finger region

**Mandatory qualifiers**
NOTE

**Comment**
The type of zinc finger is indicated in the NOTE qualifier. For example: "GATA-type" and "NR C4-type"

## SECTION 8: QUALIFIERS FOR AMINO ACID SEQUENCES

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

### 8.1. Qualifier MOL_TYPE

**Definition**
In vivo molecule type of sequence

**Value format**
protein

**Example**
<INSDQualifier_value>protein</INSDQualifier_value>

**Comment**
The "MOL_TYPE" qualifier is mandatory on the SOURCE feature key.

### 8.2. Qualifier NOTE

**Definition**
Any comment or additional information

**Value format**
free text

**Example**
<INSDQualifier_value>Heme (covalent)</INSDQualifier_value>

**Comment**
The "NOTE" qualifier is mandatory for the feature keys: BINDING; CARBOHYD; CROSSLINK; DISULFID; DNA_BIND; DOMAIN; LIPID; METAL; MOD_RES; NP_BIND and ZN_FING

### 8.3. Qualifier ORGANISM

**Definition**
Scientific name of the organism that provided the peptide

**Value format**
free text

**Example**
<INSDQualifier_value>Homo sapiens</INSDQualifier_value>

**Comment**
The "ORGANISM" qualifier is mandatory for the SOURCE feature key.
### SECTION 9: GENETIC CODES 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, 6, 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><strong>AAa</strong> = <strong>FFLLSSSSYY</strong>CC*WLLLLPPPPPHHQ**QQRRRIIIMTTTTTNNKKSSRRVVVVAAADDEEEGGGG</td>
</tr>
</tbody>
</table>
### 10 - Euplotid Nuclear Code

<table>
<thead>
<tr>
<th>AAs</th>
<th>Starts</th>
<th>Base1</th>
<th>Base2</th>
<th>Base3</th>
</tr>
</thead>
</table>
| FFLLSSSSYY**CCWLLLLPPP| M     | tttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
### 22 - Scenedesmus obliquus Mitochondrial Code

<table>
<thead>
<tr>
<th>AAs</th>
<th>Starts</th>
<th>Base1</th>
<th>Base2</th>
<th>Base3</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFLLSS<em>YY</em>LCC*WLLLLPPPPHHQQRRRIIMTTTTNNKSSRRVVVVAAAAADDEEGGGG</td>
<td>-----------------------------</td>
<td>tttttttttttttttcccccccccccccccaaaaaaagggggggggggggggg</td>
<td>ttttccccaaaggggtttttccccaaaaaggggtttttccccaaaaaggggtttttccccaaaaaggg</td>
<td>tcagtca tca tca tca tca tca tca tca tca tca tca tca tca tca tca tca</td>
</tr>
</tbody>
</table>

### 23 - Thraustochytrium Mitochondrial Code

<table>
<thead>
<tr>
<th>AAs</th>
<th>Starts</th>
<th>Base1</th>
<th>Base2</th>
<th>Base3</th>
</tr>
</thead>
<tbody>
<tr>
<td>FF<em>LSSSYY**CC</em>WLLLLPPPPHHQQRRRIIMTTTTNNKSSRRVVVVAAAAADDEEGGGG</td>
<td>-----------------------------</td>
<td>tttttttttttttttcccccccccccccccaaaaaaagggggggggggggggg</td>
<td>ttttccccaaaggggtttttccccaaaaaggggtttttccccaaaaaggggtttttccccaaaaaggg</td>
<td>tcagtca tca tca tca tca tca tca tca tca tca tca tca tca tca tca tca</td>
</tr>
</tbody>
</table>

### 24 - Pterobranchia Mitochondrial Code

<table>
<thead>
<tr>
<th>AAs</th>
<th>Starts</th>
<th>Base1</th>
<th>Base2</th>
<th>Base3</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFLLSSSY**CCWLLLLPPPPHHQQRRRIIMTTTTNNKSSSKVVVVAAAAADDEEGGGG</td>
<td>-----------------------------</td>
<td>tttttttttttttttcccccccccccccccaaaaaaagggggggggggggggg</td>
<td>ttttccccaaaggggtttttccccaaaaaggggtttttccccaaaaaggggtttttccccaaaaaggg</td>
<td>tcagtca tca tca tca tca tca tca tca tca tca tca tca tca tca tca tca</td>
</tr>
</tbody>
</table>

### 25 - Candidate Division SR1 and Gracilibacteria Code

<table>
<thead>
<tr>
<th>AAs</th>
<th>Starts</th>
<th>Base1</th>
<th>Base2</th>
<th>Base3</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFLLSSSY**CCGWLLLLPPPPHHQQRRRIIMTTTTNNKSSRRVVVVAAAAADDEEGGGG</td>
<td>-----------------------------</td>
<td>tttttttttttttttcccccccccccccccaaaaaaagggggggggggggggg</td>
<td>ttttccccaaaggggtttttccccaaaaaggggtttttccccaaaaaggggtttttccccaaaaaggg</td>
<td>tcagtca tca tca tca tca tca tca tca tca tca tca tca tca tca tca tca</td>
</tr>
</tbody>
</table>

[Annex II to ST.26 follows]
DOCUMENT TYPE DEFINITION FOR SEQUENCE LISTING (DTD)

Proposal presented by the SEQL Task Force for consideration and adoption at the CWS/4

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

This entity may be identified by the PUBLIC identifier:
********************************************************************************
PUBLIC "//WIPO//DTD SEQUENCE LISTING 1.0//EN" "ST26SequenceListing_V1_0.dtd"
********************************************************************************

************
* PUBLIC DTD URL
* http://www.wipo.int/standards/DTD/ST26SequenceListing_V1_0.dtd
********************************************************************************

Recommended Standard for the presentation of nucleotide and amino acid sequence listings using XML (eXtensible Markup Language)
********************************************************************************

* CONTACTS
********************************************************************************

xml.standards@wipo.int

Date draft created: 2014-03-11
********************************************************************************

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

* REVISION HISTORY
********************************************************************************
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) >

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

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

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

<!ELEMENT ApplicantNameLatin (#PCDATA) >

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

<!ELEMENT InventorNameLatin (#PCDATA) >

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

<!ELEMENT SequenceTotalQuantity (#PCDATA) >

<!ELEMENT SequenceData (#PCDATA) >
For intentionally skipped sequences see the ST.26 main body document.

<!--
<![CDATA[
</![CDATA[

<!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.
-->
<!ELEMENT IPOfficeCode (#PCDATA)>

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

<!--FilingDate
The date of filing of the patent application for which the sequence listing is submitted
ST.2 format (paragraphs 7 (a) and 11) “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
-->
<!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 v1.4 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.
-->
<!ELEMENT INSDSeq (INSDSeq_length,INSDSeq_moltype,INSDSeq_division,
  INSDSeq_other-seqids?,INSDSeq_feature-table?,INSDSeq_sequence)>

<!--INSDSeq_length
-->
<!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+)>

<!ELEMENT INSDSeq_sequence (INSDFeature+)>

<!ELEMENT INSDSeqid (#PCDATA)>

<!ELEMENT INSDFeature (INSDFeature_key,INSDFeature_location,INSDFeature_quals?)>

<!ELEMENT INSDFeature_key (#PCDATA)>

<!ELEMENT INSDFeature_location (#PCDATA)>

<!ELEMENT INSDFeature_quals (INSDQualifier*)>

<!ELEMENT INSDQualifier (INSDQualifier_name,INSDQualifier_value?)>

<!ELEMENT INSDQualifier_name (#PCDATA)>

<!ELEMENT INSDQualifier_value (#PCDATA)>

The residues of the sequence. The sequence must not contain numbers, punctuation or whitespace characters.

<!-- 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)

[Annex III to ST.26 follows]
ST.26 - ANNEX III  
SEQUENCE LISTING SPECIMEN (XML file)  
Final Draft

Proposal presented by the SEQL Task Force for consideration and adoption at the CWS/4
<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_sequence>GYLDVRKDVKRIKALEAFKNKMDKEK</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="3">
<INSDSeq>
<INSDSeq_length>62</INSDSeq_length>
<INSDSeq_moltype>DNA</INSDSeq_moltype>
<INSDSeq_division>PAT</INSDSeq_division>
<INSDSeq_feature-table>
<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>
<INSDFeature_key>CDS</INSDFeature_key>
<INSDFeature_location>3..62</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<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>
<INSDSeq_sequence>tgatgcctcgaccttgactgtccctcgaccccacacgcatttatagctccagcctgtgctag</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="4">
<INSDSeq>
<INSDSeq_length>19</INSDSeq_length>
<INSDSeq_moltype>AA</INSDSeq_moltype>
<INSDSeq_division>PAT</INSDSeq_division>
</INSDSeq>
</SequenceData>
<INSDSeq>
  <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>
    <INSDFeature>
      <INSDFeature_key>MOD_RES</INSDFeature_key>
      <INSDFeature_location>3</INSDFeature_location>
      <INSDFeature_quals>
        <INSDQualifier>
          <INSDQualifier_name>NOTE</INSDQualifier_name>
          <INSDQualifier_value>N-acetylalanine</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>
      </INSDFeature_quals>
    </INSDFeature>
  </INSDSeq_feature-table>
  <INSDSeq_sequence>
    atgaatattaaacanaaaagntgataaatgatatttgatatataaaaaaggttttaggttagcacagaggaaggttttagagttagcagagaaggattttgaga
cggctagtggagagacaagggcattaataaaggataaacatattgacaata
  </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>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>
</SequenceData>
<INSDQualifier_value>D-Arginine</INSDQualifier_value>
</INSDFeature>

<INSDFeature>
<INSDFeature_key>UNSURE</INSDFeature_key>
<INSDFeature_location>15</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>NONE</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>NONE</INSDQualifier_name>
<INSDQualifier_value>I, A, F, Y, all, Hie, or Nle</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>

<INSDFeature>
<INSDFeature_key>SITE</INSDFeature_key>
<INSDFeature_location>22</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>NONE</INSDQualifier_name>
<INSDQualifier_value>Homoserine</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>

<INSDSeq_sequence>GSASDVKDVKEKRIXFEXFNSKMDKSK</INSDSeq_sequence>
</INSDFeature>
</INSDSeq_sequence>
</Sequences>

<SequenceData sequenceIDNumber="7">
<INSDSeq>
<INSDSeq_length/>
<INSDSeq_moltype/>
<INSDSeq_division/>
<INSDSeq_sequence>000</INSDSeq_sequence>
</INSDSeq>
</SequenceData>

<SequenceData sequenceIDNumber="8">
<INSDSeq>
<INSDSeq_length":74</INSDSeq_length>
<INSDSeq_moltype>RNA</INSDSeq_moltype>
<INSDSeq_division>PAT</INSDSeq_division>
<INSDSeq_feature-table>
<INSDFeature>
<INSDFeature_key>source</INSDFeature_key>
<INSDFeature_location>1..74</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>organism</INSDQualifier_name>
<INSDQualifier_value>Dengue virus 2</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>mol_type</INSDQualifier_name>
<INSDQualifier_value>genomic RNA</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
</INSDSeq_sequence>
</Sequences>
<INSDSeq_sequence>atgaaattaaaacataaaagggatgataaaatgagatttgatataaaaaaggttttagagttagcagagaagga</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData_sequenceIDNumber="9">
<INSDSeq>
<INSDSeq_length>120</INSDSeq_length>
<INSDSeq_moltype>DNA</INSDSeq_moltype>
<INSDSeq_division>PAT</INSDSeq_division>
<INSDSeq_feature-table>
<INSDFeature>
<INSDFeature_key>source</INSDFeature_key>
<INSDFeature_location>1..120</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>other DNA</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
<INSDFeature>
<INSDFeature_key>misc_feature</INSDFeature_key>
<INSDFeature_location>1..60</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>note</INSDQualifier_name>
<INSDQualifier_value>DNA fragment</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
<INSDFeature>
<INSDFeature_key>misc_feature</INSDFeature_key>
<INSDFeature_location>61..120</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>note</INSDQualifier_name>
<INSDQualifier_value>RNA fragment</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>cgacccacgcgtccgaggaaccaaccatcacgtttgaggacttcgtgaaggaattggataatacccgtccctaccaaaatggcgagcgccgactcattgctcctcgtaccgtcgagcggc</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData_sequenceIDNumber="10">
<INSDSeq>
<INSDSeq_length>288</INSDSeq_length>
<INSDSeq_moltype>DNA</INSDSeq_moltype>
<INSDSeq_division>PAT</INSDSeq_division>
<INSDSeq_feature-table>
<INSDFeature>
<INSDFeature_key>source</INSDFeature_key>
<INSDFeature_location>1..288</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>organism</INSDQualifier_name>
<INSDQualifier_value>Candida albicans</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>cgcaccaogcgtcogaggaaccaacccatcacgttggagcctctggtggataggtggatctgctcctaccaaatggcgcggcgactcattgctcctggtgctctgagggc</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<INSDQualifier_name>mol_type</INSDQualifier_name>
<INSDQualifier_value>genomic DNA</INSDQualifier_value>

</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>

<INSDFeature>
<INSDFeature_key>CDS</INSDFeature_key>
<INSDFeature_location>1..288</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>translation</INSDQualifier_name>
<INSDQualifier_value>MNLTLHNVIQTDSRGEKFMKIPEIYIRGIHIKYLRIPDDIMGYAKEQSMINMENRNRYQKRRGTTSSGGGGGGGGGGGGGGGGGGGGGGGSGDSRRFNNNRQSHGNYGRR</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>transl_table</INSDQualifier_name>
<INSDQualifier_value>12</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>protein_id</INSDQualifier_name>
<INSDQualifier_value>11</INSDQualifier_value>
</INSDQualifier>

</INSDFeature_quals>
</INSDFeature>
</INSDFeature_key>
</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
<INSDQualifier_value>protein</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
</INSDFeature>
</INSDFeature>
</INSDFeature>
</INSDFeature>
</INSDFeature>
</INSDFeature>
</ST26SequenceListing>

[Annex IV to ST.26 follows]
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>SPACE</td>
<td></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>0028</td>
<td>(</td>
<td>LEFT PARENTHESIS</td>
</tr>
<tr>
<td>0029</td>
<td>)</td>
<td>RIGHT PARENTHESIS</td>
</tr>
<tr>
<td>002A</td>
<td>*</td>
<td>ASTERISK</td>
</tr>
<tr>
<td>002B</td>
<td>+</td>
<td>PLUS SIGN</td>
</tr>
<tr>
<td>002C</td>
<td>,</td>
<td>COMMA</td>
</tr>
<tr>
<td>002D</td>
<td>-</td>
<td>HYPHEN-MINUS</td>
</tr>
<tr>
<td>002E</td>
<td>.</td>
<td>FULL STOP</td>
</tr>
<tr>
<td>002F</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>003F</td>
<td>@</td>
<td>COMMERCIAL AT</td>
</tr>
<tr>
<td>0040</td>
<td>A</td>
<td>LATIN CAPITAL LETTER A</td>
</tr>
<tr>
<td>0041</td>
<td>B</td>
<td>LATIN CAPITAL LETTER B</td>
</tr>
<tr>
<td>0042</td>
<td>C</td>
<td>LATIN CAPITAL LETTER C</td>
</tr>
<tr>
<td>0043</td>
<td>D</td>
<td>LATIN CAPITAL LETTER D</td>
</tr>
<tr>
<td>0044</td>
<td>E</td>
<td>LATIN CAPITAL LETTER E</td>
</tr>
<tr>
<td>0045</td>
<td>F</td>
<td>LATIN CAPITAL LETTER F</td>
</tr>
<tr>
<td>0046</td>
<td>G</td>
<td>LATIN CAPITAL LETTER G</td>
</tr>
<tr>
<td>0047</td>
<td>H</td>
<td>LATIN CAPITAL LETTER H</td>
</tr>
<tr>
<td>0048</td>
<td>I</td>
<td>LATIN CAPITAL LETTER I</td>
</tr>
<tr>
<td>0049</td>
<td>J</td>
<td>LATIN CAPITAL LETTER J</td>
</tr>
<tr>
<td>004A</td>
<td>K</td>
<td>LATIN CAPITAL LETTER K</td>
</tr>
<tr>
<td>004B</td>
<td>L</td>
<td>LATIN CAPITAL LETTER L</td>
</tr>
<tr>
<td>004C</td>
<td>M</td>
<td>LATIN CAPITAL LETTER M</td>
</tr>
<tr>
<td>004D</td>
<td>N</td>
<td>LATIN CAPITAL LETTER N</td>
</tr>
<tr>
<td>004E</td>
<td>O</td>
<td>LATIN CAPITAL LETTER O</td>
</tr>
<tr>
<td>004F</td>
<td>P</td>
<td>LATIN CAPITAL LETTER P</td>
</tr>
<tr>
<td>0050</td>
<td>Q</td>
<td>LATIN CAPITAL LETTER Q</td>
</tr>
<tr>
<td>0051</td>
<td>R</td>
<td>LATIN CAPITAL LETTER R</td>
</tr>
<tr>
<td>0052</td>
<td>S</td>
<td>LATIN CAPITAL LETTER S</td>
</tr>
<tr>
<td>0053</td>
<td>T</td>
<td>LATIN CAPITAL LETTER T</td>
</tr>
<tr>
<td>0054</td>
<td>U</td>
<td>LATIN CAPITAL LETTER U</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>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>VERTICAL LINE</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 to ST.26 follows]
ST.26 - ANNEX V

ADDITIONAL DATA EXCHANGE REQUIREMENTS (FOR PATENT OFFICES ONLY)

Final Draft

Proposal presented by the SEQL Task Force for consideration and adoption at the CWS/4

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:

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:

pat|WO|2013999999|A1|123456

Which would be translated into a valid XML instance as:

  <INSDSeq_other-seqids>
    <INSDSeqid>pat|WO|2013999999|A1|123456</INSDSeqid>
  </INSDSeq_other-seqids>

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

[End of Annex II and of document]