STANDARD ST.26

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

Draft for Public Comment

TABLE OF CONTENTS

ANNEXES

STANDARD ST.26

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

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 represented, and the Document Type Definition (DTD) for a sequence listing in XML (eXtensible Markup Language). It is recommended that industrial property offices accept any sequence listing compliant with this Standard filed as part of a patent application or in relation to a patent application.

2. The purpose of this Standard is to:

(a) allow applicants to draw up a single sequence listing in a patent application acceptable for the purposes of both international and national or regional procedures;

(b) enhance the accuracy and quality of presentations of sequences for easier dissemination, benefiting applicants, the public and examiners;

(c) facilitate searching of the sequence data; and

(d) allow sequence data to be exchanged in electronic form and introduced into computerized databases.

DEFINITIONS

3. For the purpose of this Standard, the expression:

(a) “amino acid” means any amino acid that can be represented using any of the symbols set forth in Annex I (see Section 3, Table 3). Such amino acids include, inter alia, D-amino acids and amino acids containing modified or synthetic side chains. Amino acids will be construed as unmodified L-amino acids unless further described in the feature table as modified according to paragraph 29. For the purpose of this standard, a peptide nucleic acid (PNA) residue is not considered an amino acid, but is considered a nucleotide as set forth in paragraph 3(d)(i)B.

(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.

(b*bis*) “enumeration of its residues” means disclosure of a sequence in a patent application by listing, in order, each residue of the sequence, wherein:

(i) the residue is represented by a name, abbreviation, symbol, or structure (e.g., HHHHHHQ or HisHisHisHisHisHisGln); or

(ii) multiple residues are represented by a shorthand formula (e.g., $\mathrm{His_6Gln}$).

(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.

(cbis) “modified amino acid” means any amino acid as described in paragraph 3(a) other than L-alanine, L-arginine, L-asparagine, L-aspartic acid, L-cysteine, L-glutamine, L-glutamic acid, L-glycine, L-histidine, L-isoleucine, L-leucine, L-lysine, L-methionine, L-phenylalanine, L-proline, L-pyrrolysine, L-serine, L-selenocysteine, L-threonine, L-tryptophan, L-tyrosine, or L-valine.

(cter) “modified nucleotide” means any nucleotide as described in paragraph 3(d) other than deoxyadenosine 3’-monophosphate, deoxyguanosine 3’-monophosphate, deoxycytidine 3’-monophosphate, deoxythymidine 3’-monophosphate, adenosine 3’-monophosphate, guanosine 3’-monophosphate, cytidine 3’-monophosphate, or uridine 3’-monophosphate.

(d) "nucleotide" means any nucleotide or nucleotide analogue that can be represented using any of the symbols set forth in Annex I (see Section 1, Table 1) wherein the nucleotide or nucleotide analogue contains:

(i) a backbone moiety selected from:

A. 2' deoxyribose 5' monophosphate (the backbone moiety of a deoxyribonucleotide) or ribose 5' monophosphate (the backbone moiety of a ribonucleotide); or

B. an analogue of a 2' deoxyribose 5' monophosphate or ribose 5' monophosphate, which when forming the backbone of a nucleic acid analogue, results in an arrangement of nucleobases that mimics the arrangement of nucleobases in nucleic acids containing a 2' deoxyribose 5' monophosphate or ribose 5' monophosphate backbone, wherein the nucleic acid analogue is capable of base pairing with a complementary nucleic acid; examples of nucleotide analogues include amino acids as in peptide nucleic acids, glycol molecules as in glycol nucleic acids, threofuranosyl sugar molecules as in threose nucleic acids, morpholine rings and phosphorodiamidate groups as in morpholinos, and cyclohexenyl molecules as in cyclohexenyl nucleic acids.

and

(ii) the backbone moiety is either:

A. joined to a nucleobase, including a modified or synthetic purine or pyrimidine nucleobase; or

B. lacking a purine or pyrimidine nucleobase when the nucleotide is part of a nucleotide sequence, referred to as an "AP site" or an "abasic site".

(e) "residue" means any individual nucleotide or amino acid or their respective analogues 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 includes the disclosed nucleotide and/or amino acid sequence(s), along with any further description, as prescribed by this Standard.

(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 (see Section 1, Table 1, and Section 3, Table 3, respectively).

(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 can be represented as:

(a) an unbranched sequence or a linear region of a branched sequence containing ten or more specifically defined nucleotides, wherein adjacent nucleotides are joined by:

(i) a 3' to 5' (or 5' to 3') phosphodiester linkage; or

(ii) any chemical bond that results in an arrangement of adjacent nucleobases that mimics the arrangement of nucleobases in naturally occurring nucleic acids; or

(b) an unbranched sequence or a linear region 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/;
International Standard ISO 639-1:2002	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.

REPRESENTATION OF SEQUENCES

9. Each sequence encompassed by paragraph 6 must be assigned a separate sequence identification number, including a sequence which is identical to a region of a longer sequence. The sequence identification numbers must begin with number 1, and increase consecutively by integers. Where no sequence is present for a sequence identification number, i.e. an intentionally skipped sequence, “000” must be used in place of a sequence (see paragraph 58). The total number of sequences must be indicated in the sequence listing and must equal the total number of sequence identification numbers, whether followed by a sequence or by “000.”

Nucleotide sequences

10. A nucleotide sequence must be represented only by a single strand, in the 5’-end to 3’-end direction from left to right, or in the direction from left to right that mimics the 5’-end to 3’-end direction. The designations 5’ and 3’ or any other similar designations must not be included in the sequence. A double-stranded nucleotide sequence disclosed by enumeration of the residues of both strands must be represented 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. For the purpose of this Standard, the first nucleotide presented in the sequence is residue position number 1. When nucleotide sequences are circular in configuration, applicant must choose the nucleotide in residue position number 1. Numbering is continuous throughout the entire sequence in the direction 5’ to 3’, or in the direction that mimics the direction 5’ to 3’. The last residue position number must equal the number of nucleotides in the sequence.

12. [Removed].

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 further described in the feature table as provided by paragraph 18.

15. Where an ambiguity symbol (representing two or more alternative nucleotides) is appropriate, the most restrictive symbol should be used, as listed in Annex I (section 1, Table 1). For example, if a nucleotide 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 “t/u” 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 in the feature table, as provided in paragraphs 16 and 17 or 20. For representation of sequence variants, i.e., alternatives, deletions, insertions, or substitutions, see paragraphs 91*bis* to 97.

16. Modified nucleotides should be represented in the sequence as the corresponding unmodified nucleotides, 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), i.e., an “other” nucleotide, 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 60 *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 nucleotide must be provided as the value in a “note” qualifier. For a listing of alternative modified nucleotides, the qualifier value “OTHER” may be used in conjunction with a further “note” qualifier (see paragraphs 94 and 95). 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.

17bis. A nucleotide sequence including one or more regions of consecutive modified nucleotides that share the same backbone moiety (see paragraph 3(d)(i)B), must be further described in the feature table as provided by paragraph 17. The modified nucleotides of each such region may be jointly described in a single INSDFeature element as provided by paragraph 21. The most restrictive unabbreviated chemical name that encompasses all of the modified nucleotides in the range or a list of the chemical names of all the nucleotides in the range must be provided as the value in the “note” qualifier. For example, a glycol nucleic acid sequence containing “a”, “c”, “g”, or “t” nucleobases may be described in the “note” qualifier as “2,3-dihydroxypropyl nucleosides.” Alternatively, the same sequence may be described in the “note” qualifier as “2,3-dihydroxypropyladenine, 2,3-dihydroxypropylthymine, 2,3-dihydroxypropylguanine, or 2,3-dihydroxypropylcytosine.” Where an individual modified nucleotide in the region includes an additional modification, then the modified nucleotide must also be further described in the feature table as provided in paragraph 17.

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 representation of modified nucleotides according to paragraphs 16 to 17*bis* above:

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

```
<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 “xanthine” using “OTHER” from Annex I (see Section 2, Table 2)

```
<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>
```

Example 3: A nucleotide sequence composed of modified nucleotides encompassed by paragraph 3(d)(i)(B) with two individual nucleotides that include a further modification

```
<INSDFeature>
    <INSDFeature_key>modified_base</INSDFeature_key>
    <INSDFeature_location>1..954</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>2,3-dihydroxypropyl nucleosides</INSDQualifier_value>
            </INSDQualifier>
        </INSDFeature_quals>
    </INSDFeature>
    <INSDFeature>
        <INSDFeature_key>modified_base</INSDFeature_key>
        <INSDFeature_location>439</INSDFeature_location>
        <INSDFeature_quals>
            <INSDQualifier>
                <INSDQualifier_name>mod_base</INSDQualifier_name>
                <INSDQualifier_value>i</INSDQualifier_value>
            </INSDQualifier>
        </INSDFeature_quals>
    </INSDFeature>
    <INSDFeature>
        <INSDFeature_key>modified_base</INSDFeature_key>
        <INSDFeature_location>684</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 a single INSDFeature element with the the syntax “x..y” as the location descriptor in the element `INSDFeature_location` (see paragraphs 65 to72). For representation of sequence variants, i.e., deletions, insertions or substitutions, see paragraphs 92 to 97.

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

```
<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 an amino acid sequence must be represented in the amino to carboxy direction from left to right. The amino and carboxy groups must not be represented in the sequence.

24. For the purpose of this Standard, the first amino acid in the sequence is residue position number 1, including amino acids preceding the mature protein, for example, pre-sequences, pro-sequences, pre-pro-sequences and signal sequences. When amino acid sequences are circular in configuration, applicant must choose the amino acid in residue position number 1. Numbering is continuous through the entire 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”, “O”, “S”, “U”, “T”, “W”, “Y”, or “V”, except where it is used with a further description in the feature table as provided by paragraphs 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 in the feature table, as provided in paragraphs 28 to 30 or 31 to 33. For representation of sequence variants, i.e., alternatives, deletions, insertions, or substitutions, see paragraphs 91bis to 97.

27. Disclosed amino acid sequences separated by internal terminator symbols, represented for example by “Ter” or asterisk “*” or period “.” or a blank space, must be included 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 assigned its own sequence identification number. Terminator symbols and spaces must not be included in sequences in a sequence listing (see paragraph 57).

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

30. The following examples illustrate the representation of modified amino acids according to paragraph 29 above:

Example 1: Post-translationally modified amino acid

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

Example 2: Non post-translationally modified amino acid

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

Example 3: D-amino acid

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

31. Any “unknown” amino acid must be represented by the symbol “X” in the sequence. 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.” The symbol “X” is the equivalent of only one residue.

32. [Removed]

33. The following example illustrates the representation of an “unknown” amino acid according to paragraph 31 above:

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

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 representation 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 and assigned its own sequence identification number.

36. A sequence that contains regions of specifically defined residues separated by one or more regions of contiguous “n” or “X” residues (see paragraphs 15 and 26, respectively), wherein the exact number of “n” or “X” residues in each region is disclosed, must be included in the sequence listing as one sequence and assigned its own sequence identification number.

37. A sequence that contains regions of specifically defined residues separated by one or more gaps of an unknown or undisclosed number of residues must not be represented in the sequence listing as a single sequence. Each region of specifically defined residues that is encompassed by paragraph 6 must be included in the sequence listing as a separate sequence and assigned its own sequence identification number.

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

- o 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:

Reserved Character	Predefined Entities
<	&lt;
>	&gt;
&	&amp;
“	&quot;
'	&apos;

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:

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

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:

Element	Description	Mandatory/ Optional
`ApplicationIdentification`	The application identification for which the sequence listing is submitted	Mandatory when a sequence listing is furnished at any time following the assignment of the application number
The `ApplicationIdentification` is composed of:		
`IPOfficeCode`	ST.3 Code of the office of filing	Mandatory
`ApplicationNumberText`	The application identification as provided by the office of filing (e.g., PCT/IB2013/099999)	Mandatory
`FilingDate`	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)	Mandatory when a sequence listing is furnished at any time following the assignment of a filing date
`ApplicantFileReference`	A single unique identifier assigned by applicant to identify a particular application, typed in the characters as set forth in paragraph 40 (b)	Mandatory when a sequence listing is furnished at any time prior to assignment of the application number; otherwise, Optional
`EarliestPriorityApplicationIdentification`	The application identification of the earliest priority claim (also contains `IPOfficeCode`, `ApplicationNumberText` and `FilingDate`, see `ApplicationIdentification` above)	Mandatory where priority is claimed
`ApplicantName`	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.	Mandatory
`ApplicantNameLatin`	Where `ApplicantName` is typed in characters other than those as set forth in paragraph 40 (b), a translation or transliteration of the name of the first mentioned applicant must also be typed in characters as set forth in paragraph 40 (b)	Mandatory where `ApplicantName` contains non-Latin characters
`InventorName`	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.	Optional

Element	Description	Mandatory/ Optional
`InventorNameLatin`	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)	Optional
`InventionTitle`	Title of the invention typed in the characters as set forth in paragraph 40 (a) in the language of filing. A translation of the title of the invention into additional languages may be typed in the characters as set forth in paragraph 40 (a) using additional `InventionTitle` elements. This element includes the mandatory attribute `languageCode` as set forth in paragraph 48. The title of invention is preferably two to seven words.	Mandatory in the language of filing. Optional for additional languages.
`SequenceTotalQuantity`	The total number of all sequences in the sequence listing including intentionally skipped sequences (also known as empty sequences) (see paragraph 9).	Mandatory

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 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 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>
    <ApplicantFileReference>AB123</ApplicantFileReference>
    <EarliestPriorityApplicationIdentification>
        <IPOfficeCode>IB</IPOfficeCode>
        <ApplicationNumberText>PCT/IB2014/099999</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.
```

47. The name of the applicant and, optionally, the name of the inventor must be indicated in the element `ApplicantName` and `InventorName`, respectively, as they are generally referred to in the language in which the application is filed. The appropriate language code (see reference in paragraph 8 to ISO 639-1:2002) 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 reference in paragraph 8 to ISO 639-1:2002) 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

```
<ApplicantName languageCode="JA">出願製薬株式会社</ApplicantName>
<ApplicantNameLatin>Shutsugan Pharmaceuticals Kabushiki Kaisha</ApplicantNameLatin>
<InventorName languageCode ="JA">特許　太郎</InventorName>
<InventorNameLatin>Taro Tokkyo</InventorNameLatin>
<InventionTitle languageCode="JA"> efg タンパク質をコードするマウス 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:

Element	Description	Mandatory/Not Included	
		Sequences	Intentionally Skipped Sequences
`INSDSeq_length`	Length of the sequence	Mandatory	Mandatory with no value
`INSDSeq_moltype`	Molecule type	Mandatory	Mandatory with no value
`INSDSeq_division`	Indication that a sequence is related to a patent application	Mandatory with the value "PAT"	Mandatory with no value
`INSDSeq_feature-table`	List of annotations of the sequence	Mandatory	Must NOT be included
`INSDSeq_sequence`	Sequence	Mandatory	Mandatory with the value "000"

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 represented. For nucleotide sequences, including nucleotide analogue sequences, the molecule type must be indicated as DNA or RNA. For amino acid sequences, the molecule type must be indicated as AA. (This element is distinct from the qualifiers "mol_type" and "MOL_TYPE" discussed in paragraphs 55 and 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:

```
<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>
  cgacccacgcgtccgaggaaccaaccatcacgtttgaggacttcgtgaaggaattggataatacccgtccctaccaaaatggcg
agcgccgactcattgctcctcgtaccgtcgagcggc
    </INSDSeq_sequence>
</INSDSeq>
```

57. The element `INSDSeq_sequence` must disclose the sequence. Only the appropriate symbols set forth in Annex I (see Section 1, Table 1 and Section 3, Table 3) must be included in the sequence. The sequence must not include numbers, punctuation or whitespace characters.

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

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

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

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

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

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

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

Feature table

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

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

Element	Description	Mandatory/Optional
`INSDFeature_key`	A word or abbreviation indicating a feature	Mandatory
`INSDFeature_location`	Region of the sequence which corresponds to the feature	Mandatory
`INSDFeature_quals`	Qualifier containing auxiliary information about a feature	Mandatory where the feature key requires one or more qualifiers, e.g., source; otherwise, Optional

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. [Removed]

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 `INSDSeq_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 `INSDSeq_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 `INSDSeq_sequence` element, and x is less than y.

Location descriptor type	Syntax	Description
Single residue number	`x`	Points to a single residue in the sequence.
Residue numbers delimitating a sequence span	`x..y`	Points to a continuous range of residues bounded by and including the starting and ending residues.

Residues before the first or beyond the last specified residue number	`<x`
	`>x`
	`<x..y`
	`x..>y`
		Points to a region including a specified residue or span of residues and extending beyond a specified residue. The '<' and '>' symbols may be used with a single residue or the starting and ending residue numbers of a span of residues to indicate that a features extends beyond the specified residue number.

A site between two adjoining residue numbers	`x^y`	Points to a site between two adjoining residues, e.g. endonucleolytic cleavage site. The position numbers for the adjacent residues are separated by a carat (^). The permitted formats for this descriptor are x^x+1 (for example 55^56), or, for circular nucleotides, x^1, where “x” is the full length of the molecule, i.e. 1000^1 for circular molecule with length 1000.

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:

Location syntax	Location description
`join(location,location, ... location)`	The indicated locations are joined (placed end-to-end) to form one contiguous sequence.
`order(location,location, ... location)`	The elements are found in the specified order but nothing is implied about whether joining those elements is reasonable.

(b) Location operator for nucleotides only:

Location syntax	Location description
`complement(location)`	Indicates that the feature is located on the strand complementary to the sequence span specified by the location descriptor, when read in the 5’ to 3’ direction or in the direction that mimics the 5’ to 3’ direction.

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:

Location Example	Description
`467`	Points to residue 467 in the sequence.
`123^124`	Points to a site between residues 123 and 124.
`340..565`	Points to a continuous range of residues bounded by and including residues 340 and 565.
`<1`	Points to a feature location before the first residue.

Location Example	Description
`<345..500`	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.
`<1..888`	Indicates that the feature starts before the first sequence residue and continues to and includes residue 888.
`1..>888`	Indicates that the feature starts at the first sequenced residue and continues beyond residue 888.
`join(12..78,134..202)`	Indicates that regions 12 to 78 and 134 to 202 should be joined to form one contiguous sequence.

(b) locations for nucleotides only:

Location example	Description
`complement(34..126)`	Start at the nucleotide complementary to 126 and finish at the nucleotide complementary to nucleotide 34 (the feature is on the strand complementary to the presented strand).
`complement(join(2691..4571, 4918..5163))`	Joins nucleotides 2691 to 4571 and 4918 to 5163, then complements the joined segments (the feature is on the strand complementary to the presented strand).
`join(complement(4918..5163), complement(2691..4571))`	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).

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 included in the sequence listing and assigned 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:

Element	Description	Mandatory/Optional
`INSDQualifier_name`	Name of the qualifier (see Annex I, Sections 6 and 8)	Mandatory
`INSDQualifier_value`	Value of the qualifier, if any, in the specified format (see Annex I, Sections 6 and 8)	Mandatory, when specified (see Annex I, Sections 6 and 8)

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. 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 a sequence as per paragraphs 78 and 79 above:

Example 1: Source for a nucleotide sequence

```
<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 an amino acid sequence

```
<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:

```
<INSDQualifier_name>organism</INSDQualifier_name>
<INSDQualifier_value>Bacillus sp.</INSDQualifier_value>
```

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

```
<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 `INSDSeq_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 `INSDFeature_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.

88bis The “transl_except” qualifier must be used with the “CDS” feature key and the “translation” qualifier to identify a codon that encodes either pyrrolysine or selenocysteine.

90. An amino acid sequence encoded by the coding sequence and disclosed in a “translation” qualifier that is encompassed by paragraph 6 must be included in the sequence listing and assigned its own sequence identification number. The sequence identification number assigned to the amino acid sequence must be provided as the value in the qualifier “protein_id” with the “CDS” feature key. The “ORGANISM” qualifier of the “SOURCE” feature key for the amino acid sequence must be identical to that of its coding sequence. For example:

```
<INSDSeq_feature-table>
    <INSDFeature>
        <INSDFeature_key>CDS</INSDFeature_key>
        <INSDFeature_location>1..507</INSDFeature_location>
        <INSDFeature_quals>
            <INSDQualifier>
                <INSDQualifier_name>transl_table</INSDQualifier_name>
                <INSDQualifier_value>11</INSDQualifier_value>
            </INSDQualifier>
            <INSDQualifier>
                <INSDQualifier_name>translation</INSDQualifier_name>
                <INSDQualifier_value>
MLVHLERTTIMFDFSSLINLPLIWGLLIAIAVLLYILMDGFDLGIGILLPFAPSDKCRDHMISSIAPFWDGNETWLVLGGGGLFAA
FPLAYSILMPAFYIPIIIMLLGLIVRGVSFEFRFKAEGKYRRLWDYAFHFGSLGAAFCQGMILGAFIHGVEVNGRNFSGGQLM
                </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 each be included in the sequence listing and assigned their own sequence identification number.

91bis Any variant sequence, disclosed as a single sequence with enumerated alternative variant residues at one or more positions, must be included in the sequence listing and should be represented by a single sequence, wherein the enumerated alternative variant residues are represented by the most restrictive ambiguity symbol (see paragraphs 15 and 26).

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

(a) may be represented 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 represented as a separate sequence and assigned 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 represented as a separate sequence and assigned 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:

Type of sequence	Feature Key	Qualifier	Use
Nucleic acid	variation	replace or note	Naturally occurring mutations and polymorphisms, e.g. alleles, RFLPs.
Nucleic acid	misc_difference	replace or note	Variability introduced artificially, e.g. by genetic manipulation or by chemical synthesis.

Type of sequence	Feature Key	Qualifier	Use
Amino acid	VAR_SEQ	NOTE	Variant produced by alternative splicing, alternative promoter usage, alternative initiation and ribosomal frameshifting.
Amino acid	VARIANT	NOTE	Any type of variant for which VAR_SEQ is not applicable.

94. Annotation of a sequence for a specific variant must include a feature key and qualifier, as indicated in the table above, and the feature location. The value for the “replace” qualifier must be only a single alternative nucleotide or nucleotide sequence using only the symbols in set forth Section 1, Table 1. A listing of alternative variant residues may be provided as the value in the “note” or “NOTE” qualifier. In particular, a listing of alternative amino acids must be provided as the value in the “NOTE” qualifier where “X” is used in a sequence, but represents a subgroup of “any one of ‘A’, ‘R’, ‘N’, ‘D’, ‘C’, ‘Q’, ‘E’, ‘G’, ‘H’, ‘I’, ‘L’, ‘K’, ‘M’, ‘F’, ‘P’, ‘O’, ‘S’, ‘U’, ‘T’, ‘W’, ‘Y’, or ‘V’”. A deletion must be represented by an empty qualifier value for the qualifier “replace” or an indication in the “note” or “NOTE” that the residue may be deleted. An inserted or substituted residue(s) must be provided in the “replace”, “note”, or “NOTE” qualifier. The value format for the “replace”, “note”, and “NOTE” qualifiers is free text and must not exceed 1000 characters, as provided in paragraph 87. See paragraph 97 for sequences encompassed by paragraph 6 that are provided as an insertion or a substitution in a 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. For the “note” or “NOTE” qualifier, where the variant residue is a modified residue not set forth in Tables 2 or 4 of Annex I, the complete unabbreviated name of the modified residue must be provided as the qualifier value. Modified residues must be further described in the feature table as provided in paragraph 17 or 29

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

Example 1: Feature key “misc_difference” for enumerated alternative variant nucleotides.
The “n” at position 53 of the sequence can be one of five alternative nucleotides.

```
<INSDFeature>
    <INSDFeature_key>misc_difference</INSDFeature_key>
    <INSDFeature_location>53</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>w, cmnm5s2u, mam5u, mcm5s2u, or
p</INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
</INSDFeature>
<INSDFeature>
    <INSDFeature_key>modified_base</INSDFeature_key>
    <INSDFeature_location>53</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>mod_base</INSDQualifier_name>
            <INSDQualifier_value>OTHER</INSDQualifier_value>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>cmnm5s2u, mam5u, mcm5s2u, or p</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 “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 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 `I`, `A`, `F`, `Y`, `aIle`, `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>
<INSDFeature>
    <INSDFeature_key>MOD_RES</INSDFeature_key>
    <INSDFeature_location>100</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>aIle, MeIle, or Nle</INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
</INSDFeature>
```

Example 6: Feature key “VARIANT” for a substitution in an amino acid sequence.
The amino acid given in position 100 of the sequence can be replaced by any amino acid except for Lys, Arg or His.

```
<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 included in the sequence listing and assigned its own sequence identification number.

[Annex I to ST.26 follows]