**ISTANDARD ST.26**

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

*Version 1.6*

*Proposal presented for consideration by the Committee on WIPO Standards (CWS)*

*at its tenth session*

*Editorial Note prepared by the International Bureau*

*The Tenth Session of the Committee on WIPO Standards decided that version 1.6 of WIPO Standard ST.26 will enter into force on July 1, 2023. Meanwhile, version 1.5 of Standard ST.26 should continue to be used.*

*.*

*The Standard is published for information purposes of intellectual property offices and other interested parties.*

TABLE OF CONTENTS

[INTRODUCTION 3](#_Toc90370558)

[DEFINITIONS 3](#_Toc90370559)

[SCOPE 4](#_Toc90370560)

[REFERENCES 5](#_Toc90370561)

[REPRESENTATION OF SEQUENCES 5](#_Toc90370562)

[*Nucleotide sequences* 5](#_Toc90370563)

[*Amino acid sequences* 8](#_Toc90370564)

[*Presentation of special situations* 9](#_Toc90370565)

[STRUCTURE OF THE SEQUENCE LISTING IN XML 9](#_Toc90370566)

[*Root element* 10](#_Toc90370567)

[*General information part* 11](#_Toc90370568)

[*Sequence data part* 14](#_Toc90370569)

[*Feature table* 16](#_Toc90370570)

[*Feature keys* 17](#_Toc90370571)

[*Mandatory feature keys* 17](#_Toc90370572)

[*Feature location* 17](#_Toc90370573)

[*Feature qualifiers* 19](#_Toc90370574)

[*Mandatory feature qualifiers* 19](#_Toc90370575)

[*Qualifier elements* 19](#_Toc90370576)

[*Free text* 21](#_Toc90370577)

[*Coding sequences* 23](#_Toc90370578)

[*Variants* 23](#_Toc90370579)

**ANNEXES**

[Annex I](#Annex1)  - Controlled vocabulary

[Annex II](#_ANNEX_II)  - Document Type Definition (DTD) for Sequence Listing

[Annex III](#_ANNEX_III)  - Sequence Listing Specimen (XML file)

[Annex IV](#_ANNEX_IV)  - Character Subset from the Unicode Basic Latin Code Table for Use in an XML Instance of a Sequence Listing

[Annex V](#_ANNEX_V)  - Additional data exchange requirements (for IPOs only)

[Annex VI](#_ANNEX_VI)  - Guidance document with illustrated examples

[Appendix](#AnnexVIAppendix) - Guidance document sequences in XML

[Annex VII](#_ANNEX_VII) - Recommendation for the transformation of a sequence listing from ST.25 to ST.26:

potential added or deleted subject matter

**STANDARD ST.26**

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

*Version 1.6*

*Proposal presented for consideration by the Committee on WIPO Standards (CWS)*

*at its tenth session*

## INTRODUCTION

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 intellectual property offices accept any sequence listing compliant with this Standard filed as part of a patent application or in relation to a patent application.

The purpose of this Standard is to:

1. 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;
2. enhance the accuracy and quality of presentations of sequences for easier dissemination, benefiting applicants, the public and examiners;
3. facilitate searching of the sequence data; and
4. allow sequence data to be exchanged in electronic form and introduced into computerized databases.

## DEFINITIONS

For the purpose of this Standard, the expression:

1. “amino acid” means any amino acid that can be represented using any of the symbols set forth in Annex I (see Section 3, Table 3). Such amino acids include, inter alia, D-amino acids and amino acids containing modified or synthetic side chains. Amino acids will be construed as unmodified L-amino acids unless further described in the feature table as modified according to paragraph 30. For the purpose of this standard, a peptide nucleic acid (PNA) residue is not considered an amino acid, but is considered a nucleotide as set forth in paragraph 3(g)(i)(2).
2. “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.
3. “enumeration of its residues” means disclosure of a sequence in a patent application by listing, in order, each residue of the sequence, wherein:
4. the residue is represented by a name, abbreviation, symbol, or structure (e.g., HHHHHHQ or HisHisHisHisHisHisGln); or
5. multiple residues are represented by a shorthand formula (e.g., His6Gln).
6. “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.
7. “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.
8. “modified nucleotide” means any nucleotide as described in paragraph 3(g) 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.
9. “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:

* + - 1. 2’ deoxyribose 5’ monophosphate (the backbone moiety of a deoxyribonucleotide) or ribose 5’ monophosphate (the backbone moiety of a ribonucleotide); or
      2. 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 backbone moieties 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:

* + - 1. joined to a nucleobase, including a modified or synthetic purine or pyrimidine nucleobase; or
      2. 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”.

1. “residue” means any individual nucleotide or amino acid or their respective analogues in a sequence.
2. “sequence identification number” means a unique number (integer) assigned to each sequence in the sequence listing.
3. “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.
4. “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).
5. “unknown” nucleotide or amino acid means that a single nucleotide or amino acid is present but its identity is unknown or not disclosed.
6. “variant sequence” means a nucleotide or amino acid sequence that contains one or more differences with respect to a primary sequence. These differences may include alternative residues (see paragraphs 15 and 27), modified residues (see paragraphs 3(g), 3(h), 16, and 29), deletions, insertions, and substitutions. See paragraphs 93 to 95.
7. “free text” is a type of value format for certain qualifiers, presented in the form of a descriptive text phrase or other specified format (as indicated in Annex I). See paragraph 85.
8. “language-dependent free text” means the free text value of certain qualifiers, which may require translation for international, national or regional procedures. See paragraph 87.

For the purpose of this Standard, the word(s):

(a) “may” refers to an optional or permissible approach, but not a requirement.

(b) “must” refers to a requirement of the Standard; disregard of the requirement will result in noncompliance.

(c) “must not” refers to a prohibition of the Standard.

(d) “should” refers to a strongly encouraged approach, but not a requirement.

(e) “should not” refers to a strongly discouraged approach, but not a prohibition.

## SCOPE

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

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.

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:

1. 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

1. an unbranched sequence or a linear region of a branched sequence containing four or more specifically defined amino acids, wherein the amino acids form a single peptide backbone, i.e. adjacent amino acids are joined by peptide bonds.

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

## REFERENCES

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](http://www.wipo.int/standards/en/pdf/03-02-01.pdf) Standard manner for designating calendar dates by using Gregorian calendar;

WIPO Standard [ST.3](http://www.wipo.int/standards/en/pdf/03-03-01.pdf) Recommended standard on two-letter codes for the representation of states, other entities and intergovernmental organizations;

WIPO Standard [ST.16](http://www.wipo.int/standards/en/pdf/03-16-01.pdf) Recommended standard code for the identification of different kinds of patent documents;

WIPO Standard [ST.25](http://www.wipo.int/standards/en/pdf/03-25-01.pdf) Standard for the presentation of nucleotide and amino acid sequence listings in patent applications.

## REPRESENTATION OF SEQUENCES

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

### *Nucleotide sequences*

A nucleotide sequence must be represented only by a single strand, in the 5’ to 3’ direction from left to right, or in the direction from left to right that mimics the 5’ to 3’ 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:

1. 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
2. two separate sequences, each assigned its own sequence identification number, where the two strands are not fully complementary to each other.

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 5’ to 3’ direction, or in the direction that mimics the 5’ to 3’ direction. The last residue position number must equal the number of nucleotides in the sequence.

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.

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 19.

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 in the feature table. The symbol “n” must 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, 17, 21, or 93-96. For representation of sequence variants, i.e., alternatives, deletions, insertions, or substitutions, see paragraphs 93 to 100.

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 a non-naturally occurring nucleotide, must be represented by the symbol “n”. The symbol “n” is the equivalent of only one residue.

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 97 and 98). 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.

A nucleotide sequence including one or more regions of consecutive modified nucleotides that share the same backbone moiety (see paragraph 3(g)(i)(2)), 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 22. 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.

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.

The following examples illustrate the representation of modified nucleotides according to paragraphs 16 to 18 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 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(g)(i)(2) 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>

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.

A region containing a known number of contiguous “a”, “c”, “g”, “t”, or “n” residues for which the same description applies may be jointly described using a single INSDFeature element with the syntax “x..y” as the location descriptor in the element INSDFeature\_location (see paragraphs 64 to 71). For representation of sequence variants, i.e., alternatives, deletions, insertions or substitutions, see paragraphs 93 to 100.

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

<INSDFeature>

<INSDFeature\_key>modified\_base</INSDFeature\_key>

<INSDFeature\_location>358..485</INSDFeature\_location>

<INSDFeature\_quals>

<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*

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.

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 an amino acid sequence is circular in configuration and the ring consists solely of amino acid residues linked by peptide bonds, i.e., the sequence has no amino and carboxy termini, applicant must choose the amino acid in residue position number 1. Numbering is continuous through the entire sequence in the amino to carboxy direction.

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.

Where an ambiguity symbol (representing two or more amino acids in the alternative) is appropriate, the most restrictive symbol should be used, as listed in Annex I (Section 3, Table 3). 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. The symbol “X” must not be used to represent anything other than an amino acid. A single modified or “unknown” amino acid may be represented by the symbol “X”, together with a further description in the feature table, e.g., as provided by paragraphs 29, 30, 32, or 93-98. For representation of sequence variants, i.e., alternatives, deletions, insertions, or substitutions, see paragraphs 93 to 100.

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 7. 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).

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.

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.

The following examples illustrate the representation of modified amino acids according to paragraph 30 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>

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.

The following example illustrates the representation of an “unknown” amino acid according to paragraph 32 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>

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 64 to 70). For representation of sequence variants, i.e., alternatives, deletions, insertions, or substitutions, see paragraphs 93 to 100.

### *Presentation of special situations*

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.

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 27, 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.

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 7 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

In accordance with paragraph 6 above, an XML instance of a sequence listing file according to this Standard is composed of:

1. a general information part, which contains information concerning the patent application to which the sequence listing is directed; and
2. a 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.

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

1. The first line of the XML instance must contain the XML declaration:

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

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

<!DOCTYPE ST26SequenceListing PUBLIC "-//WIPO//DTD Sequence Listing 1.3//EN" "ST26SequenceListing\_V1\_3.dtd">.

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

1. the information contained in the elements ApplicantName, InventorName and InventionTitle of the general information part, and the NonEnglishQualifier\_value of the sequence data part, may be composed of any valid Unicode characters indicated in the XML 1.0 specification except the Unicode Control code points 0000-001F and 007F-009F. The reserved characters “, &, ‘, <, and > (Unicode code points 0022, 0026, 0027, 003C and 003E respectively), must be replaced as set forth in paragraph 41; and
2. the information contained in all other elements and attributes of the general information part and in all other elements and attributes of the sequence data part must be composed of printable characters (including the space character) from the Unicode Basic Latin code table (i.e., limited to Unicode code points 0020 through 007E – see Annex IV). The reserved characters “, &, ‘, <, and > (Unicode code points 0022, 0026, 0027, 003C and 003E respectively), must be replaced as set forth in paragraph 41.

In an XML instance of a sequence listing, numeric character references[[1]](#footnote-2) must not be used and 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 71 for an example. The only character entity references permitted are the predefined entities set forth in this paragraph.

All mandatory elements must be populated (except as provided for in paragraph 58 for an intentionally skipped sequence). Optional elements for which content is not available should not appear in the XML instance (except as provided for in paragraph 97 for representation of a deletion in a sequence in the value for the qualifier “replace”).

### *Root element*

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\_3”. | 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 |
| originalFreeTextLanguageCode | The language code (see reference in paragraph 9 to ISO 639-1:2002) for the single original language in which the language-dependent free text qualifiers were prepared. | Optional |
| nonEnglishFreeTextLanguageCode | The language code (see reference in paragraph 9 to ISO 639-1:2002) for the NonEnglishQualifier\_value elements | Mandatory when a NonEnglishQualifier\_value element is present in the sequence listing |

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

<ST26SequenceListing dtdVersion="V1\_3" fileName="US11-405455-SEQL.xml" softwareName="WIPO Sequence" softwareVersion="1.0" productionDate="2022-05-10" originalFreeTextLanguageCode="de" nonEnglishFreeTextLanguageCode="fr">

{...}\*

</ST26SequenceListing>

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

### *General information part*

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

| **Element** | **Description** | **Mandatory/**  **Optional** |
| --- | --- | --- |
| ApplicationIdentification  The ApplicationIdentification is composed of: | 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 |
| IPOfficeCode | ST.3 Code of the office of filing | Mandatory |
| ApplicationNumberText | The  application number 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 identification of the earliest priority application (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 |
| 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 should be between 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 10). | Mandatory |

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 1.3//EN" "ST26SequenceListing\_V1\_3.dtd">

<ST26SequenceListing dtdVersion="V1\_3" fileName="Invention\_SEQL.xml" softwareName="WIPO Sequence" softwareVersion="1.0" productionDate="2022-05-10" originalFreeTextLanguageCode="en" nonEnglishFreeTextLanguageCode="ja">

<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 1.3//EN" "ST26SequenceListing\_V1\_3.dtd">

<ST26SequenceListing dtdVersion="1\_3" fileName="Invention\_SEQL.xml" softwareName="WIPO Sequence" softwareVersion="1.0" productionDate="2022-05-10" originalFreeTextLanguageCode="en" nonEnglishFreeTextLanguageCode="ja">

<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.

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 9 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.

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 9 to ISO 639-1:2002) must be indicated in the languageCode attribute of the element.

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*

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

Each SequenceData element must have a mandatory attribute sequenceIDNumber, in which the sequence identification number (see paragraph 10) for each sequence is contained. For example:

<SequenceData sequenceIDNumber="1">

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” |

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>

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 qualifier “mol\_type” discussed in paragraphs 55 and 84). For example:

<INSDSeq\_moltype>AA</INSDSeq\_moltype>

For a nucleotide sequence that contains both DNA and RNA segments of one or more nucleotides, the molecule type must be indicated as 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 segment of the combined DNA/RNA molecule must be further described with the feature key “misc\_feature” and the qualifier “note”, which indicates whether the segment is DNA or RNA.

The following example illustrates the description of a nucleotide sequence containing both DNA and RNA segments 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</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</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

</INSDSeq\_feature-table> <INSDSeq\_sequence>cgacccacgcgtccgaggaaccaaccatcacgtttgaggacttcgtgaaggaattggataatacccgtccctaccaaaatggcgagcgccgactcattgctcctcgtaccgtcgagcggc</INSDSeq\_sequence>

</INSDSeq>

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.

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

1. the element SequenceData and its attribute sequenceIDNumber, with the sequence identification number of the skipped sequence provided as the value;
2. the elements INSDSeq\_length, INSDSeq\_moltype, INSDSeq\_division, present but with no value provided;
3. the element INSDSeq\_feature-table must not be included; and
4. the element INSDSeq\_sequence with the string “000” as the value.

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*

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.

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*

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*

The “source” feature key is mandatory for all nucleotide sequences and for all amino acid sequences, except for any intentionally skipped sequence. Each sequence must have a single “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.

### *Feature location*

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. Amino acid sequences must contain one and only one location descriptor in the mandatory INSDFeature\_location element. Nucleotide sequences may have more than one location descriptor in the mandatory INSDFeature\_location element when used in conjunction with one or more location operator(s) (see paragraphs 67 to 70).

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

The syntax for each type of location descriptor is indicated in the table below, where x and y are residue numbers, indicated as positive integers, not greater than the length of the sequence in the INSDSeq\_sequence element, and x is less than y.

1. Location descriptors for nucleotide and amino acid sequences:

| **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  <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 feature extends beyond the specified residue number. |

1. Location descriptors for nucleotide sequences only:

| **Location descriptor type** | **Syntax** | **Description** |
| --- | --- | --- |
| A site between two adjoining nucleotides | x^y | Points to a site between two adjoining nucleotides, e.g., endonucleolytic cleavage site. The position numbers for the adjacent nucleotides 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. |

1. Location descriptors for amino acid sequences only:

| **Location descriptor type** | **Syntax** | **Description** |
| --- | --- | --- |
| Residue numbers joined by an intrachain cross-link | x..y | Points to amino acids joined by an intrachain linkage when used with a feature that indicates an intrachain cross-link, such as “CROSSLNK” or “DISULFID”. |

The INSDFeature\_location element of nucleotide sequences may contain one or more location operators. 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. Location operators can be used for nucleotides only.

|  |  |
| --- | --- |
| **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. |
| 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. |

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

The location operator “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.

The following examples illustrate feature locations, as per paragraphs 64 to 69 above:

1. locations for nucleotide and amino acid sequences:

| **Location Example** | **Description** |
| --- | --- |
| 467 | Points to residue 467 in the sequence. |
| 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. |
| <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 sequenced 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. |
| <1..>888 | Indicates that the feature starts before the first sequenced residue and continues beyond residue 888. |

1. locations for nucleotide sequences only:

|  |  |
| --- | --- |
| **Location example** | **Description** |
| 123^124 | Points to a site between residues 123 and 124. |
| join(12..78,134..202) | Indicates that regions 12 to 78 and 134 to 202 should be joined to form one contiguous sequence. |
| complement(34..126) | Starts at the nucleotide complementary to 126 and finishes 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). |

(c) locations for amino acid sequences only:

|  |  |
| --- | --- |
| **Location example** | **Description** |
| 340..565 | Indicates that the amino acids at positions 340 and 565 are joined by an intrachain linkage when used with a feature that indicates an intrachain cross-link, such as “CROSSLNK” or “DISULFID”. |

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*

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:

1. free text (see paragraphs 85 to 87);
2. controlled vocabulary or enumerated values (e.g., a number or date); and
3. sequences.

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

Any sequence encompassed by paragraph 7 which is provided as a qualifier value must be separately included in the sequence listing and assigned its own sequence identification number (see paragraph 10).

### *Mandatory feature qualifiers*

One mandatory feature key, i.e., “source” for nucleotide sequences and amino acid sequences, requires two mandatory qualifiers, “organism” and “mol\_type”. Some optional feature keys also require mandatory qualifiers.

### *Qualifier elements*

The element INSDFeature\_quals contains one or more INSDQualifier elements. Each INSDQualifier element represents a single qualifier and consists of three dependent elements and one attribute as follows:

|  |  |  |
| --- | --- | --- |
| **Element/Attribute** | **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) and  composed in the characters as set forth in paragraph 40(b). | Mandatory, when specified (see paragraph 87 and Annex I, Sections 6 and 8) |
| NonEnglishQualifier\_value | Value of the qualifier, if any, in the specified format (see Annex I, Sections 6 and 8) and composed in the characters as set forth in paragraph 40(a). | Mandatory, when specified (see paragraph 87 and Annex I, Sections 6 and 8) |
| id | A qualifier with language-dependent free text value may be uniquely identified by using the optional XML attribute 'id' in the element INSDQualifier (see paragraph 87(d)). The value of the 'id' attribute must start with the letter 'q' and continue with any positive integer. The value of an 'id' attribute must be unique to one INSDQualifier element, i.e. the attribute value must only be used once in a sequence listing file. | Optional |

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.

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 amino acid sequences, but must not be used in the organism qualifier value.

The following examples illustrate the source of a sequence as per paragraphs 77 and 78 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>

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>

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”. Any known taxonomic information should be indicated 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>

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>

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>

The “mol\_type” qualifier for nucleotide sequences (see Annex I, Section 6) and “mol\_type” qualifier 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:

1. 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”;
2. For an amino acid sequences, the “mol\_type” qualifier value is “protein”.

### *Free text*

Free text, as indicated in paragraph 3 (n), is a type of value format for certain qualifiers presented in the form of a descriptive text phrase or other specified format (as indicated in Annex I).

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 other than the "translation" qualifier, the free text must not exceed 1000 characters.

Language-dependent free text, as indicated in paragraph 3 (o), is the free text value of certain qualifiers that is language-dependent in that it may require translation for international, national or regional procedures. Qualifiers for nucleotide sequences with a language-dependent free text value format are identified in Annex I, Section 6, Table 5. Qualifiers for amino acid sequences with a language-dependent free text value format are identified in Annex I, Section 8, Table 6.

(a) Language-dependent free text must be presented in the INSDQualifier\_value element in English, or in the NonEnglishQualifier\_value element in a language other than English, or in both elements. Note that if an organism name is a Latin genus and species name, no translation is required. Technical terms and proper names originating from non-English words that are used internationally are considered English for the purpose of the value of the INSDQualifier\_value element (e.g., ‘in vitro’, ‘in vivo’).

(b) If a NonEnglishQualifier\_value element is present in a sequence listing, the appropriate language code (see reference in paragraph 9 to ISO 639-1:2002) must be indicated in the nonEnglishFreeTextLanguageCode attribute in the root element (see paragraph 43). All NonEnglishQualifier\_value elements in a single sequence listing must have values in the language indicated in the nonEnglishFreeTextLanguageCode attribute. The NonEnglishQualifier\_value element is permitted only for qualifiers that have a language-dependent free text value format.

(c) Where NonEnglishQualifier\_value and INSDQualifier\_value are both present for a single qualifier, the information contained in the two elements must be equivalent. One of the following conditions must be true: NonEnglishQualifier\_value contains a translation of the value of INSDQualifier\_value; or, INSDQualifier\_value contains a translation of the value of NonEnglishQualifier\_value; or, both elements contain a translation of the qualifier value from the language specified in the originalFreeTextLanguageCode attribute (see paragraph 43).

(d) For qualifiers with a language-dependent free text value,the INSDQualifier element may include an optional attribute id. The value of this attribute must be in the format "q" followed by a positive integer, e.g. "q23", and must be unique to one INSDQualifier element, i.e. the attribute value must only be used once in a sequence listing file.

The following examples illustrate the presentation of language-dependent free text as discussed in paragraph 87.

Example 1: language-dependent free text in an INSDQualifier\_value element:

<INSDFeature>

<INSDFeature\_key>regulatory</INSDFeature\_key>

<INSDFeature\_location>1..60</INSDFeature\_location>

<INSDFeature\_quals>

<INSDQualifier id="q1">

<INSDQualifier\_name>function</INSDQualifier\_name>

<INSDQualifier\_value>binds to regulatory protein Est3</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

Example 2: language-dependent free text in an INSDQualifier\_value element and a NonEnglishQualifier\_value element:

<INSDFeature>

<INSDFeature\_key>ACT\_SITE</INSDFeature\_key>

<INSDFeature\_location>51..64</INSDFeature\_location>

<INSDFeature\_quals>

<INSDQualifier id="q45">

<INSDQualifier\_name>note</INSDQualifier\_name>

<INSDQualifier\_value>cleaves carbohydrate chain</INSDQualifier\_value>

<NonEnglishQualifier\_value>clive la chaîne glucidique

</NonEnglishQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

Example 3: language-dependent free text in a NonEnglishQualifier\_value element:

<INSDFeature>

<INSDFeature\_key>ACT\_SITE</INSDFeature\_key>

<INSDFeature\_location>51..64</INSDFeature\_location>

<INSDFeature\_quals>

<INSDQualifier id="q1034">

<INSDQualifier\_name>note</INSDQualifier\_name>

<NonEnglishQualifier\_value>clive la chaîne glucidique

</NonEnglishQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

### *Coding sequences*

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 location of the “CDS” feature in the mandatory element INSDFeature\_location must include the stop codon.

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 7) is assumed.

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.

An amino acid sequence encoded by the coding sequence and disclosed in a “translation” qualifier that is encompassed by paragraph 7 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:

<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>MLVHLERTTIMFDFSSLINLPLIWGLLIAIAVLLYILMDGFDLGIGILLPFAPSDKCRDHMISSIAPFWDGNETWLVLGGGGLFAAFPLAYSILMPAFYIPIIIMLLGLIVRGVSFEFRFKAEGKYRRLWDYAFHFGSLGAAFCQGMILGAFIHGVEVNGRNFSGGQLM

</INSDQualifier\_value>

</INSDQualifier>

<INSDQualifier>

<INSDQualifier\_name>protein\_id</INSDQualifier\_name>

<INSDQualifier\_value>89</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

### *Variants*

A primary sequence and any variant of that sequence, each disclosed by enumeration of their residues and encompassed by paragraph 7, must each be included in the sequence listing and assigned their own sequence identification number.

Any variant sequence, disclosed as a single sequence with enumerated alternative 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 residues are represented by the most restrictive ambiguity symbol (see paragraphs 15 and 27).

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

1. 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;
2. 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
3. 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 86).

The table below indicates the proper use of feature keys and qualifiers for nucleic acid and amino acid sequence 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. |
| 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. |

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, or empty. A listing of alternative residues may be provided as the value in the “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, and represents a value other than “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’” (see paragraph 27). A deletion must be represented by an empty qualifier value for the “replace” qualifier or by an indication in the “note” qualifier that the residue may be deleted. 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 86. See paragraph 100 for sequences encompassed by paragraph 7 that are provided as an insertion or a substitution in a qualifier value.

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” 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 30.

The following examples illustrate the representation of variants as per paragraphs 95 to 98 above:

Example 1: Feature key “misc\_difference” for enumerated alternative 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>

<INSDQualifier>

<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>

A sequence encompassed by paragraph 7 that is provided as an insertion or a substitution in a qualifier value for a primary sequence annotation must also be included in the sequence listing and assigned its own sequence identification number.

[Annex I follows]

## **ANNEX I**

CONTROLLED VOCABULARY

*Version 1.6*

*Proposal presented for consideration by the Committee on WIPO Standards (CWS)*

*at its tenth session*

TABLE OF CONTENTS

[SECTION 1: LIST OF NUCLEOTIDES 2](#_Toc530474320)

[SECTION 2: LIST OF MODIFIED NUCLEOTIDES 2](#_Toc530474321)

[SECTION 3: LIST OF AMINO ACIDS 4](#_Toc530474322)

[SECTION 4: LIST OF MODIFIED AMINO ACIDS 5](#_Toc530474323)

[SECTION 5: FEATURE KEYS FOR NUCLEOTIDE SEQUENCES 6](#_Toc530474324)

[SECTION 6: QUALIFIERS FOR NUCLEOTIDE SEQUENCES 24](#_Toc530474374)

[SECTION 7: FEATURE KEYS FOR AMINO ACID SEQUENCES 46](#_Toc530474455)

[SECTION 8: QUALIFIERS FOR AMINO ACID SEQUENCES 53](#_Toc530474495)

[SECTION 9: GENETIC CODE TABLES 54](#_Toc530474499)

## SECTION 1: LIST OF NUCLEOTIDES

The nucleotide base symbols 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 symbols

|  |  |
| --- | --- |
| **Symbol** | **Definition** |
| a | adenine |
| c | cytosine |
| g | guanine |
| t | thymine in DNA/uracil in RNA (t/u) |
| m | a or c |
| r | a or g |
| w | a or t/u |
| s | c or g |
| y | c or t/u |
| k | g or t/u |
| v | a or c or g; not t/u |
| h | a or c or t/u; not g |
| d | a or g or t/u; not c |
| b | c or g or t/u; not a |
| n | a or c or g or t/u; “unknown” or “other” |

## 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

| **Abbreviation** | **Definition** |
| --- | --- |
| ac4c | 4-acetylcytidine |
| chm5u | 5-(carboxyhydroxylmethyl)uridine |
| cm | 2'-O-methylcytidine |
| cmnm5s2u | 5-carboxymethylaminomethyl-2-thiouridine |
| cmnm5u | 5-carboxymethylaminomethyluridine |
| dhu | dihydrouridine |
| fm | 2'-O-methylpseudouridine |
| gal q | beta-D-galactosylqueuosine |
| gm | 2'-O-methylguanosine |
| i | inosine |
| i6a | N6-isopentenyladenosine |
| m1a | 1-methyladenosine |
| m1f | 1-methylpseudouridine |
| m1g | 1-methylguanosine |
| m1i | 1-methylinosine |
| m22g | 2,2-dimethylguanosine |
| m2a | 2-methyladenosine |
| m2g | 2-methylguanosine |
| m3c | 3-methylcytidine |
| m4c | N4-methylcytosine |
| m5c | 5-methylcytidine |
| m6a | N6-methyladenosine |
| m7g | 7-methylguanosine |
| mam5u | 5-methylaminomethyluridine |
| mam5s2u | 5-methylaminomethyl-2-thiouridine |
| man q | beta-D-mannosylqueuosine |
| mcm5s2u | 5-methoxycarbonylmethyl-2-thiouridine |
| mcm5u | 5-methoxycarbonylmethyluridine |
| mo5u | 5-methoxyuridine |
| ms2i6a | 2-methylthio-N6-isopentenyladenosine |
| ms2t6a | N-((9-beta-D-ribofuranosyl-2-methylthiopurine-6-yl)carbamoyl)threonine |
| mt6a | N-((9-beta-D-ribofuranosylpurine-6-yl)N-methyl-carbamoyl)threonine |
| mv | uridine-5-oxoacetic acid-methylester |
| o5u | uridine-5-oxyacetic acid (v) |
| osyw | wybutoxosine |
| p | pseudouridine |
| q | queuosine |
| s2c | 2-thiocytidine |
| s2t | 5-methyl-2-thiouridine |
| s2u | 2-thiouridine |
| s4u | 4-thiouridine |
| m5u | 5-methyluridine |
| t6a | N-((9-beta-D-ribofuranosylpurine-6-yl)carbamoyl)threonine |
| tm | 2'-O-methyl-5-methyluridine |
| um | 2'-O-methyluridine |
| yw | wybutosine |
| x | 3-(3-amino-3-carboxypropyl)uridine, (acp3)u |
| OTHER | (requires note qualifier) |

## SECTION 3: LIST OF AMINO ACIDS

The amino acid symbols 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 symbols

|  |  |
| --- | --- |
| **Symbol** | **Definition** |
| A | Alanine |
| R | Arginine |
| N | Asparagine |
| D | Aspartic acid (Aspartate) |
| C | Cysteine |
| Q | Glutamine |
| E | Glutamic acid (Glutamate) |
| G | Glycine |
| H | Histidine |
| I | Isoleucine |
| L | Leucine |
| K | Lysine |
| M | Methionine |
| F | Phenylalanine |
| P | Proline |
| O | Pyrrolysine |
| S | Serine |
| U | Selenocysteine |
| T | Threonine |
| W | Tryptophan |
| Y | Tyrosine |
| V | Valine |
| B | Aspartic acid or Asparagine |
| Z | Glutamine or Glutamic acid |
| J | Leucine or Isoleucine |
| X | A or R or N or D or C or Q or E or G or H or I or L or K or M or F or P or O or S or U or T or W or Y or V; “unknown” or “other” |

## SECTION 4: LIST OF MODIFIED AMINO ACIDS

Table 4 lists the only permitted abbreviations for a modified 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 amino acids

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

## SECTION 5: FEATURE KEYS FOR NUCLEOTIDE SEQUENCES

This section 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”. 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.

* 1. 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

Organism scope eukaryotes

* 1. 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

circular\_RNA

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, pseugogene or pseudo are permitted with a CDS feature key; when the translation qualifier is used, the protein\_id qualifier is mandatory if the translation product contains four or more specifically defined amino acids

* 1. 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

* 1. 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

* 1. Feature Key D\_segment

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

Optional qualifiers allele

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

Organism scope eukaryotes

* 1. 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

* 1. 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.

* 1. Feature Key iDNA

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

Optional qualifiers allele

function

gene

gene\_synonym

map

note

number

standard\_name

Molecule scope DNA

Comment e.g., in the somatic processing of immunoglobulin genes.

* 1. Feature Key intron

Definition a segment of DNA that is transcribed, but removed from within the transcript by splicing together the sequences (exons) on either side of it

Optional qualifiers allele

function

gene

gene\_synonym

map

note

number

pseudo

pseudogene

standard\_name

trans\_splicing

* 1. Feature Key J\_segment

Definition joining segment of immunoglobulin light and heavy

chains, and T-cell receptor alpha, beta, and gamma chains

Optional qualifiers allele

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

Organism scope eukaryotes

* 1. 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

* 1. Feature Key misc\_binding

Definition site in nucleic acid which covalently or non-covalently binds another moiety that cannot be described by any other binding key (primer\_bind or protein\_bind)

Mandatory qualifiers bound\_moiety

Optional qualifiers allele

function

gene

gene\_synonym

map

note

Comment note that the regulatory feature key and regulatory\_class qualifier with the value “ribosome\_binding\_site” must be used for describing ribosome binding sites

* 1. Feature Key misc\_difference

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

Optional qualifiers allele

clone

compare

gene

gene\_synonym

map

note

phenotype

replace

standard\_name

Comment the misc\_difference feature key must be used to describe variability introduced artificially, e.g., by genetic manipulation or by chemical synthesis; use the replace qualifier to annotate a deletion, insertion, or substitution. The variation feature key must be used to describe naturally occurring genetic variability.

* 1. 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

* 1. Feature Key misc\_recomb

Definition site of any generalized, site-specific or replicative recombination event where there is a breakage and reunion of duplex DNA that cannot be described by other recombination keys or qualifiers of source key (proviral)

Optional qualifiers allele

gene

gene\_synonym

map

note

recombination\_class

standard\_name

Molecule scope DNA

* 1. 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

* 1. Feature Key misc\_structure

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

Optional qualifiers allele

function

gene

gene\_synonym

map

note

standard\_name

* 1. Feature Key mobile\_element

Definition region of genome containing mobile elements

Mandatory qualifiers mobile\_element\_type

Optional qualifiers allele

function

gene

gene\_synonym

map

note

rpt\_family

rpt\_type

standard\_name

* 1. Feature Key modified\_base

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

Mandatory qualifiers mod\_base

Optional qualifiers allele

frequency

gene

gene\_synonym

map

note

Comment value for the mandatory mod\_base qualifier is limited to the restricted vocabulary for modified base abbreviations in Section 2 of this Annex.

* 1. 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

circular\_RNA

function

gene

gene\_synonym

map

note

operon

product

pseudo

pseudogene

standard\_name

trans\_splicing

* 1. 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 must not be used for ribosomal and transfer RNA annotation, for which the rRNA and tRNA feature keys must be used, respectively

* 1. 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

Organism scope eukaryotes

* 1. Feature Key operon

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

Mandatory qualifiers operon

Optional qualifiers allele

function

map

note

phenotype

pseudo

pseudogene

standard\_name

* 1. Feature Key oriT

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

Optional qualifiers allele

bound\_moiety

direction

gene

gene\_synonym

map

note

rpt\_family

rpt\_type

rpt\_unit\_range

rpt\_unit\_seq

standard\_name

Molecule Scope DNA

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

* 1. Feature Key polyA\_site

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

Optional qualifiers allele

gene

gene\_synonym

map

note

Organism scope eukaryotes and eukaryotic viruses

* 1. Feature Key precursor\_RNA

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

Optional qualifiers allele

function

gene

gene\_synonym

map

note

operon

product

standard\_name

trans\_splicing

Comment used for RNA which may be the result of post-transcriptional processing; if the RNA in question is known not to have been processed, use the prim\_transcript key

* 1. Feature Key prim\_transcript

Definition primary (initial, unprocessed) transcript; may include ncRNA, rRNA, tRNA, 5’ untranslated region (5’UTR), coding sequences (CDS, exon), intervening sequences (intron) and 3’ untranslated region (3’UTR)

Optional qualifiers allele

function

gene

gene\_synonym

map

note

operon

standard\_name

* 1. 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

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; 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

* 1. Feature Key propeptide

Definition propeptide coding sequence; coding sequence for the domain of a proprotein that is cleaved to form the mature protein product.

Optional qualifiers allele

function

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

* 1. 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 the regulatory feature key and regulatory\_class qualifier with the value “ribosome\_binding\_site” must be used to describe ribosome binding sites

* 1. Feature Key regulatory

Definition any region of a sequence that functions in the regulation of transcription, translation, replication or chromatin structure;

Mandatory qualifiers regulatory\_class

Optional qualifiers allele

bound\_moiety

function

gene

gene\_synonym

map

note

operon

phenotype

pseudo

pseudogene

standard\_name

* 1. 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

* 1. Feature Key rep\_origin

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

Optional Qualifiers allele

direction

function

gene

gene\_synonym

map

note

standard\_name

Comment direction qualifier has valid values: left, right, or both

* 1. Feature Key rRNA

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

Optional qualifiers allele

function

gene

gene\_synonym

map

note

operon

product

pseudo

pseudogene

standard\_name

Comment rRNA sizes should be annotated with the product qualifier

* 1. Feature Key S\_region

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

Optional qualifiers allele

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

Organism scope eukaryotes

* 1. Feature Key sig\_peptide

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

Optional qualifiers allele

function

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

* 1. 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

* 1. 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

* 1. Feature Key STS

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

Optional qualifiers allele

gene

gene\_synonym

map

note

standard\_name

Molecule scope DNA

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

* 1. Feature Key telomere

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

Optional qualifiers note

rpt\_type

rpt\_unit\_range

rpt\_unit\_seq

standard\_name

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

* 1. Feature Key tmRNA

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

Optional qualifiers allele

function

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

tag\_peptide

* 1. 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

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

* 1. 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

circular\_RNA

anticodon

function

gene

gene\_synonym

map

note

operon

product

pseudo

pseudogene

standard\_name

trans\_splicing

* 1. Feature Key unsure

Definition a small region of sequenced bases, generally 10 or fewer in its length, which could not be confidently identified. Such a region might contain called bases (a, t, g, or c), or a mixture of called-bases and uncalled-bases ('n').

Optional qualifiers allele

compare

gene

gene\_synonym

map

note

replace

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

* 1. Feature Key V\_region

Definition variable region of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; codes for the variable amino terminal portion; can be composed of V\_segments, D\_segments, N\_regions, and J\_segments

Optional qualifiers allele

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

Organism scope eukaryotes

* 1. Feature Key V\_segment

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

Optional qualifiers allele

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

Organism scope eukaryotes

* 1. 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, RFLP’s, and other naturally occurring mutations and polymorphisms; use the replace qualifier to annotate a deletion, insertion, or substitution; variability arising as a result of genetic manipulation (e.g., site directed mutagenesis) must be described with the misc\_difference feature

* 1. Feature Key 3’UTR

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

Optional qualifiers allele

function

gene

gene\_synonym

map

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&apos;UTR” in the XML file, i.e., <INSDFeature\_key>3&apos;UTR</INSDFeature\_key>.

* 1. Feature Key 5’UTR

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

Optional qualifiers allele

function

gene

gene\_synonym

map

note

standard\_name

trans\_splicing

Comment The apostrophe character has special meaning in XML, and must be substituted with “&apos;” in the value of an element. Thus “5’UTR” must be represented as “5&apos;UTR” in the XML file, i.e., <INSDFeature\_key>5&apos;UTR</INSDFeature\_key>.

## 

## SECTION 6: QUALIFIERS FOR NUCLEOTIDE SEQUENCES

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

Where the value format is “none”, the INSDQualifier\_value element must not be used and the NonEnglishQualifier\_value element must not be used.

Where the value format is free text that is identified as language-dependent, one of the following must be used:

1) the INSDQualifier\_value element; or

2) the NonEnglishQualifier\_value element; or

3) both the INSDQualifier\_value element and the NonEnglishQualifier\_value element.

Where the value format is something other than “none” but not identified as language-dependent free text, the INSDQualifier\_value element must be used and the NonEnglishQualifier\_value element must not be used.

PLEASE NOTE: Any qualifier value provided for a qualifier with a language-dependent free text value format may require translation for international, national or, regional procedures. The qualifiers listed in the following table are considered to have language-dependent free text values:

Table 5: List of qualifiers for nucleotide sequences with language-dependent free text values

|  |  |
| --- | --- |
| **Section** | **Language-Dependent Free Text Qualifier** |
| 6.3 | bound\_moiety |
| 6.5 | cell\_type |
| 6.8 | clone |
| 6.9 | clone\_lib |
| 6.11 | collected\_by |
| 6.14 | cultivar |
| 6.15 | dev\_stage |
| 6.18 | ecotype |
| 6.21 | frequency |
| 6.22 | function |
| 6.24 | gene\_synonym |
| 6.26 | haplogroup |
| 6.28 | host |
| 6.29 | identified\_by |
| 6.30 | isolate |
| 6.31 | isolation\_source |
| 6.32 | lab\_host |
| 6.36 | mating\_type |
| 6.41 | note |
| 6.45 | organism |
| 6.47 | phenotype |
| 6.49 | pop\_variant |
| 6.50 | product |
| 6.66 | serotype |
| 6.67 | serovar |
| 6.68 | sex |
| 6.69 | standard\_name |
| 6.70 | strain |
| 6.71 | sub\_clone |
| 6.72 | sub\_species |
| 6.73 | sub\_strain |
| 6.75 | tissue\_lib |
| 6.76 | tissue\_type |
| 6.81 | variety |

* 1. Qualifier allele

Definition name of the allele for the given gene

Mandatory 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.

* 1. Qualifier anticodon

Definition location of the anticodon of tRNA and the amino acid for which it codes

Mandatory value format (pos:<location>,aa:<amino\_acid>,seq:<text>) where <location> is the position of the anticodon and <amino\_acid> is the three letter abbreviation for the amino acid encoded and <text> 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>

* 1. Qualifier bound\_moiety

Definition name of the molecule/complex that may bind to the given feature

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>GAL4</INSDQualifier\_value>

Comment A single bound\_moiety qualifier is permitted on the "misc\_binding", "oriT" and "protein\_bind" features.

* 1. Qualifier cell\_line

Definition cell line from which the sequence was obtained

Mandatory value format free text

Example <INSDQualifier\_value>MCF7</INSDQualifier\_value>

* 1. Qualifier cell\_type

Definition cell type from which the sequence was obtained

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>leukocyte</INSDQualifier\_value>

* 1. Qualifier chromosome

Definition chromosome (e.g., Chromosome number) from which the sequence was obtained

Mandatory value format free text

Example <INSDQualifier\_value>1</INSDQualifier\_value>

<INSDQualifier\_value>X</INSDQualifier\_value>

* 1. Qualifier circular\_RNA

Definition indicates that exons are out-of-order or overlapping because this spliced RNA product is a circular RNA (circRNA) created by backsplicing, for example when a downstream exon in the gene is located 5' of an upstream exon in the RNA product

Value format none

Comment should be used on features such as CDS, mRNA, tRNA and other features that are produced as a result of a backsplicing event. This qualifier should be used only when the splice event is indicated in the "join" operator, eg join(complement(69611..69724),139856..140087)

* 1. Qualifier clone

Definition clone from which the sequence was obtained

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>lambda-hIL7.3</INSDQualifier\_value>

Comment a source feature must not contain more than one clone qualifier; where the sequence was obtained from multiple clones it may be further described in the feature table using the feature key misc\_feature and a note qualifier to specify the multiple clones.

* 1. Qualifier clone\_lib

Definition clone library from which the sequence was obtained

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>lambda-hIL7</INSDQualifier\_value>

* 1. 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.

Mandatory value format 1 or 2 or 3

Example <INSDQualifier\_value>2</INSDQualifier\_value>

* 1. Qualifier collected\_by

Definition name of persons or institute who collected the specimen

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>Dan Janzen</INSDQualifier\_value>

* 1. Qualifier collection\_date

Definition date that the specimen was collected.

Mandatory value format YYYY-MM-DD, YYYY-MM or YYYY

Example <INSDQualifier\_value>1952-10-21</INSDQualifier\_value>

<INSDQualifier\_value>1952-10</INSDQualifier\_value>

<INSDQualifier\_value>1952</INSDQualifier\_value>

Comment 'YYYY' is a four-digit value representing the year. 'MM' is a two-digit value representing the month. 'DD' is a two-digit value representing the day of the month.

* 1. Qualifier compare

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

Mandatory 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.

* 1. Qualifier cultivar

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

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>Nipponbare</INSDQualifier\_value>

<INSDQualifier\_value>Tenuifolius</INSDQualifier\_value>

<INSDQualifier\_value>Candy Cane</INSDQualifier\_value>

<INSDQualifier\_value>IR36</INSDQualifier\_value>

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

* 1. Qualifier dev\_stage

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

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>fourth instar larva</INSDQualifier\_value>

* 1. Qualifier direction

Definition direction of DNA replication

Mandatory value format left, right, or both  
where left indicates toward the 5’ end of the sequence (as presented) and right indicates toward the 3’ end

Example <INSDQualifier\_value>left</INSDQualifier\_value>

Comment The values left, right, and both are permitted when the direction qualifier is used to annotate a rep\_origin feature key. However, only left and right values are permitted when the direction qualifier is used to annotate an oriT feature key.

* 1. Qualifier EC\_number

Definition Enzyme Commission number for enzyme product of sequence

Mandatory 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>

<INSDQualifier\_value>1.1.2.n1</INSDQualifier\_value>

Comment valid values for EC numbers are defined in the list prepared by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) (published in Enzyme Nomenclature 1992, Academic Press, San Diego, or a more recent revision thereof).The format represents a string of four numbers separated by full stops; up to three numbers starting from the end of the string may be replaced by dash "-" to indicate uncertain assignment. Symbols including an "n", e.g., “n”, “n1” and so on, may be used in the last position instead of a number where the EC number is awaiting assignment. Please note that such incomplete EC numbers are not approved by NC-IUBMB.

* 1. Qualifier ecotype

Definition a population within a given species displaying genetically based, phenotypic traits that reflect adaptation to a local habitat

Mandatory value Format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

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.

* 1. Qualifier environmental\_sample

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

Value format none

Comment used only with the source feature key; source feature keys containing the environmental\_sample qualifier should also contain the isolation\_source qualifier; a source feature including the environmental\_sample qualifier must not include the strain qualifier.

* 1. Qualifier exception

Definition indicates that the coding region cannot be translated using standard biological rules

Mandatory value format One of the following controlled vocabulary phrases:

RNA editing

rearrangement required for product

annotated by transcript or proteomic data

Example <INSDQualifier\_value>RNA editing</INSDQualifier\_value>

<INSDQualifier\_value>rearrangement required for product</INSDQualifier\_value>

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

* 1. Qualifier frequency

Definition frequency of the occurrence of a feature

Mandatory value format free text representing the proportion of a population carrying the feature expressed as a fraction

Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>23/108</INSDQualifier\_value>

<INSDQualifier\_value>1 in 12</INSDQualifier\_value>

<INSDQualifier\_value>0.85</INSDQualifier\_value>

* 1. Qualifier function

Definition function attributed to a sequence

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>essential for recognition of cofactor </INSDQualifier\_value>

Comment The function qualifier is used when the gene name and/or product name do not convey the function attributable to a sequence.

* 1. Qualifier gene

Definition symbol of the gene corresponding to a sequence region

Mandatory 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.

* 1. Qualifier gene\_synonym

Definition synonymous, replaced, obsolete or former gene symbol

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

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 the gene\_synonym qualifier is used, a primary gene symbol must always be indicated in a gene qualifier

* 1. 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 must not be used to indicate that the source of the sequence is a gamete or germ cell; germline and rearranged qualifiers must not be used in the same source feature; germline and rearranged qualifiers must only be used for molecules that can undergo somatic rearrangements as part of an adaptive immune response; these are the T-cell receptor (TCR) and immunoglobulin loci in the jawed vertebrates, and the unrelated variable lymphocyte receptor (VLR) locus in the jawless fish (lampreys and hagfish); germline and rearranged qualifiers should not be used outside of the Craniata (taxid=89593)

* 1. 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.

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>H\*</INSDQualifier\_value>

* 1. 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.

Mandatory value format free text

Example <INSDQualifier\_value>Dw3 B5 Cw1 A1</INSDQualifier\_value>

* 1. Qualifier host

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

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>Homo sapiens</INSDQualifier\_value>

<INSDQualifier\_value>Homo sapiens 12 year old girl</INSDQualifier\_value>

<INSDQualifier\_value>Rhizobium NGR234</INSDQualifier\_value>

* 1. Qualifier identified\_by

Definition name of the expert who identified the specimen taxonomically

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>John Burns</INSDQualifier\_value>

* 1. Qualifier isolate

Definition individual isolate from which the sequence was obtained

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>Patient #152</INSDQualifier\_value>

<INSDQualifier\_value>DGGE band PSBAC-13</INSDQualifier\_value>

* 1. Qualifier isolation\_source

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

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

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

* 1. Qualifier lab\_host

Definition scientific name of the laboratory host used to propagate the source organism from which the sequenced molecule was obtained

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>Gallus gallus</INSDQualifier\_value>

<INSDQualifier\_value>Gallus gallus embryo</INSDQualifier\_value>

<INSDQualifier\_value>Escherichia coli strain DH5 alpha</INSDQualifier\_value>

<INSDQualifier\_value>Homo sapiens HeLa cells</INSDQualifier\_value>

Comment the full binomial scientific name of the host organism should be used when known; extra conditional information relating to the host may also be included

* 1. Qualifier lat\_lon

Definition geographical coordinates of the location where the specimen was collected

Mandatory value format free text - degrees latitude and longitude in format "d[d.dddd] N|S d[dd.dddd] W|E"

Example <INSDQualifier\_value>47.94 N 28.12 W</INSDQualifier\_value>

<INSDQualifier\_value>45.0123 S 4.1234 E</INSDQualifier\_value>

* 1. 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

* 1. Qualifier map

Definition genomic map position of feature

Mandatory value format free text

Example <INSDQualifier\_value>8q12-q13</INSDQualifier\_value>

* 1. 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

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Examples <INSDQualifier\_value>MAT-1</INSDQualifier\_value>

<INSDQualifier\_value>plus</INSDQualifier\_value>

<INSDQualifier\_value>-</INSDQualifier\_value>

<INSDQualifier\_value>odd</INSDQualifier\_value>

<INSDQualifier\_value>even</INSDQualifier\_value>"

Comment mating\_type qualifier values male and female are valid in the prokaryotes, but not in the eukaryotes; for more information, see the entry for the sex qualifier.

* 1. Qualifier mobile\_element\_type

Definition type and name or identifier of the mobile element which is described by the parent feature

Mandatory value format <mobile\_element\_type>[:<mobile\_element\_name>]   
where <mobile\_element\_type> is one of the following:

transposon

retrotransposon

integron

insertion sequence

non-LTR retrotransposon

SINE

MITE

LINE

other

Example <INSDQualifier\_value>transposon:Tnp9</INSDQualifier\_value>

Comment mobile\_element\_type is permitted 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>

* 1. Qualifier mod\_base

Definition abbreviation for a modified nucleotide base

Mandatory value format modified base abbreviation chosen from this Annex, Section 2

Example <INSDQualifier\_value>m5c</INSDQualifier\_value>

<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

* 1. Qualifier mol\_type

Definition molecule type of sequence

Mandatory value format One chosen from the following:

genomic DNA

genomic RNA

mRNA

tRNA

rRNA

other RNA

other DNA

transcribed RNA

viral cRNA

unassigned DNA

unassigned RNA

Example <INSDQualifier\_value>genomic DNA</INSDQualifier\_value>

<INSDQualifier\_value>other RNA</INSDQualifier\_value>

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

* 1. Qualifier ncRNA\_class

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

Mandatory value format TYPE  
where TYPE is one of the following controlled vocabulary terms or phrases:

antisense\_RNA

autocatalytically\_spliced\_intron

circRNA

ribozyme

hammerhead\_ribozyme

lncRNA

RNase\_P\_RNA

RNase\_MRP\_RNA

telomerase\_RNA

guide\_RNA

sgRNA

rasiRNA

scRNA

scaRNA

siRNA

pre\_miRNA

miRNA

piRNA

snoRNA

snRNA

SRP\_RNA

vault\_RNA

Y\_RNA

other

Example <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 must be annotated by entering "other" as the ncRNA\_class qualifier value, and providing a brief explanation of novel ncRNA\_class in a note qualifier

* 1. Qualifier note

Definition any comment or additional information

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>A comment about the feature</INSDQualifier\_value>

* 1. Qualifier number

Definition a number to indicate the order of genetic elements (e.g., exons or introns) in the 5’ to 3’ direction

Mandatory value format free text (with no whitespace characters)

Example <INSDQualifier\_value>4</INSDQualifier\_value>

<INSDQualifier\_value>6B</INSDQualifier\_value>

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

* 1. Qualifier operon

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

Mandatory value format free text

Example <INSDQualifier\_value>lac</INSDQualifier\_value>

* 1. Qualifier organelle

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

Mandatory 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>

* 1. Qualifier organism

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

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>Homo sapiens</INSDQualifier\_value>

* 1. 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

Mandatory value format [fwd\_name: XXX1, ]fwd\_seq: xxxxx1,[fwd\_name: XXX2, ]fwd\_seq: xxxxx2, [rev\_name: YYY1, ]rev\_seq: yyyyy1,[rev\_name: YYY2, ]rev\_seq: yyyyy2

Example <INSDQualifier\_value>fwd\_name: CO1P1, fwd\_seq: ttgattttttggtcayccwgaagt,rev\_name: CO1R4, rev\_seq: ccwvytardcctarraartgttg</INSDQualifier\_value>

<INSDQualifier\_value>fwd\_name: hoge1, fwd\_seq: cgkgtgtatcttact, rev\_name: hoge2, rev\_seq: cg&lt;i&gt;gtgtatcttact</INSDQualifier\_value>

<INSDQualifier\_value>fwd\_name: CO1P1, fwd\_seq: ttgattttttggtcayccwgaagt, fwd\_name: CO1P2, fwd\_seq: gatacacaggtcayccwgaagt, rev\_name: CO1R4, rev\_seq: ccwvytardcctarraartgttg</INSDQualifier\_value>

Comment fwd\_seq and rev\_seq are both mandatory; fwd\_name and rev\_name are both optional. Both sequences must be presented in 5’ to 3’ order. The sequences must be given in the symbols from Section 1 of this Annex, except for the modified bases, which must be enclosed within angle brackets < >. In XML, the angle brackets < and > must be substituted with &lt; and &gt; since they are reserved characters in XML.

* 1. Qualifier phenotype

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

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>erythromycin resistance</INSDQualifier\_value>

* 1. 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

Mandatory value format free text

Example <INSDQualifier\_value>pC589</INSDQualifier\_value>

* 1. Qualifier pop\_variant

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

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>pop1</INSDQualifier\_value>

<INSDQualifier\_value>Bear Paw</INSDQualifier\_value>

* 1. 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.

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

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)

* 1. Qualifier protein\_id

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

Mandatory value format an integer greater than zero

Example <INSDQualifier\_value>89</INSDQualifier\_value>

* 1. Qualifier proviral

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

Value format none

* 1. Qualifier pseudo

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

Value format none

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

* 1. Qualifier pseudogene

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

Mandatory value format TYPE

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

Example <INSDQualifier\_value>processed</INSDQualifier\_value>

<INSDQualifier\_value>unprocessed</INSDQualifier\_value>

<INSDQualifier\_value>unitary</INSDQualifier\_value>

<INSDQualifier\_value>allelic</INSDQualifier\_value>

<INSDQualifier\_value>unknown</INSDQualifier\_value>

Comment Definitions of TYPE values:

processed - the pseudogene has arisen by reverse transcription of a mRNA into cDNA, followed by reintegration into the genome. Therefore, it has lost any intron/exon structure, and it might have a pseudo-polyA-tail.

unprocessed - the pseudogene has arisen from a copy of the parent gene by duplication followed by accumulation of random mutations. The changes, compared to their functional homolog, include insertions, deletions, premature stop codons, frameshifts and a higher proportion of non-synonymous versus synonymous substitutions.

unitary - the pseudogene has no parent. It is the original gene, which is functional is some species but disrupted in some way (indels, mutation, recombination) in another species or strain.

allelic - a (unitary) pseudogene that is stable in the population but importantly it has a functional alternative allele also in the population. i.e., one strain may have the gene, another strain may have the pseudogene. MHC haplotypes have allelic pseudogenes.

unknown - the submitter does not know the method of pseudogenization.

* 1. Qualifier rearranged

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

Value format none

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

* 1. Qualifier recombination\_class

Definition a structured description of the classification of recombination hotspot region within a sequence

Mandatory value format TYPE  
where TYPE is one of the following controlled vocabulary terms or phrases:

meiotic

mitotic

non\_allelic\_homologous

chromosome\_breakpoint

other

Example <INSDQualifier\_value>meiotic</INSDQualifier\_value>

<INSDQualifier\_value>chromosome\_breakpoint</INSDQualifier\_value>

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

* 1. Qualifier regulatory\_class

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

Mandatory value format TYPE  
where TYPE is one of the following controlled vocabulary terms or phrases:

attenuator

CAAT\_signal

DNase\_I\_hypersensitive\_site

enhancer

enhancer\_blocking\_element

GC\_signal

imprinting\_control\_region

insulator

locus\_control\_region

matrix\_attachment\_region

minus\_35\_signal

minus\_10\_signal

polyA\_signal\_sequence

promoter

recoding\_stimulatory\_region

recombination\_enhancer

replication\_regulatory\_region

response\_element

ribosome\_binding\_site

riboswitch

silencer

TATA\_box

terminator

transcriptional\_cis\_regulatory\_region

uORF

other

Example <INSDQualifier\_value>promoter</INSDQualifier\_value>

<INSDQualifier\_value>enhancer</INSDQualifier\_value>

<INSDQualifier\_value>ribosome\_binding\_site</INSDQualifier\_value>

Comment specific regulatory classes not yet in the regulatory\_class controlled vocabulary must be annotated by entering “other” as the regulatory\_class qualifier value and providing a brief explanation of the novel regulatory\_class in a note qualifier

* 1. 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

Mandatory value format free text

Example <INSDQualifier\_value>a</INSDQualifier\_value>

<INSDQualifier\_value></INSDQualifier\_value> - for a deletion

* 1. Qualifier ribosomal\_slippage

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

Value format none

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

* 1. Qualifier rpt\_family

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

Mandatory value format free text

Example <INSDQualifier\_value>Alu</INSDQualifier\_value>

* 1. Qualifier rpt\_type

Definition structure and distribution of repeated sequence

Mandatory value format One of the following controlled vocabulary terms or phrases:

tandem

direct

inverted

flanking

nested

terminal

dispersed

long\_terminal\_repeat

non\_ltr\_retrotransposon\_polymeric\_tract

centromeric\_repeat

telomeric\_repeat

x\_element\_combinatorial\_repeat

y\_prime\_element

other

Example <INSDQualifier\_value>inverted</INSDQualifier\_value>  
<INSDQualifier\_value>long\_terminal\_repeat</INSDQualifier\_value>

Comment Definitions of the values:

tandem - a repeat that exists adjacent to another in the same orientation;

direct - a repeat that exists not always adjacent but is in the same orientation;

inverted – a repeat pair occurring in reverse orientation to one another on the same molecule;

flanking - a repeat lying outside the sequence for which it has functional significance (eg. transposon insertion target sites);

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

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

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

long\_terminal\_repeat - a sequence directly repeated at both ends of a defined sequence, of the sort typically found in retroviruses;

non\_ltr\_retrotransposon\_polymeric\_tract - a polymeric tract, such as poly(dA), within a non LTR retrotransposon;

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

telomeric\_repeat - a repeat region found within the telomere;

x\_element\_combinatorial\_repeat - a repeat region located between the X element and the telomere or adjacent Y' element;

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

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

* 1. Qualifier rpt\_unit\_range

Definition location of a repeating unit expressed as a range

Mandatory 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.

* 1. Qualifier rpt\_unit\_seq

Definition identity of a repeat sequence

Mandatory 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

* 1. Qualifier satellite

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

Mandatory 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.

* 1. Qualifier segment

Definition name of viral or phage segment sequenced

Mandatory value format free text

Example <INSDQualifier\_value>6</INSDQualifier\_value>

* 1. Qualifier serotype

Definition serological variety of a species characterized by its antigenic properties

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

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".

* 1. Qualifier serovar

Definition serological variety of a species (usually a prokaryote) characterized by its antigenic properties

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

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".

* 1. 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

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

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

* 1. Qualifier standard\_name

Definition accepted standard name for this feature

Mandatory value format free text  
this value may require translation for International/National/Regional procedures

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).

* 1. Qualifier strain

Definition strain from which sequence was obtained

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>BALB/c</INSDQualifier\_value>

Comment feature entries including a strain qualifier must not include the environmental\_sample qualifier

* 1. Qualifier sub\_clone

Definition sub-clone from which sequence was obtained

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>lambda-hIL7.20g</INSDQualifier\_value>

Comment a source feature must not contain more than one sub\_clone qualifier; to indicate that the sequence was obtained from multiple sub\_clones, multiple sources may be further described using the feature key “misc\_feature” and the qualifier “note”

* 1. Qualifier sub\_species

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

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>lactis</INSDQualifier\_value>

* 1. 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

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>abis</INSDQualifier\_value>

Comment must be accompanied by a strain qualifier in a source feature; if the parental strain is not given, the modified strain 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

* 1. Qualifier tag\_peptide

Definition base location encoding the polypeptide for proteolysis tag of tmRNA and its termination codon

Mandatory 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

* 1. Qualifier tissue\_lib

Definition tissue library from which sequence was obtained

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>tissue library 772</INSDQualifier\_value>

* 1. Qualifier tissue\_type

Definition tissue type from which the sequence was obtained

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>liver</INSDQualifier\_value>

* 1. Qualifier transl\_except

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

Mandatory value format (pos:<location>,aa:<amino\_acid>) where <amino\_acid> is the three letter abbreviation for the amino acid coded by the codon at the base\_range position

Example <INSDQualifier\_value>(pos:213..215,aa:Trp)</INSDQualifier\_value>

<INSDQualifier\_value>(pos:462..464,aa:OTHER)</INSDQualifier\_value>

<INSDQualifier\_value>(pos:1017,aa:TERM)</INSDQualifier\_value>

<INSDQualifier\_value>(pos:2000..2001,aa:TERM)</INSDQualifier\_value>

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

* 1. 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

Mandatory 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.

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

* 1. 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

Mandatory 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>MASTFPPWYRGCASTPSLKGLIMCTW</INSDQualifier\_value>

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

* 1. Qualifier variety

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

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

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.

* 1. Feature Key ACT\_SITE

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

Optional qualifiers note

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

* 1. 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.

* 1. Feature Key CA\_BIND

Definition Extent of a calcium-binding region

Optional qualifiers note

* 1. Feature Key CARBOHYD

Definition Glycosylation site

Mandatory qualifiers note

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

* 1. Feature Key CHAIN

Definition Extent of a polypeptide chain in the mature protein

Optional qualifiers note

* 1. Feature Key COILED

Definition Extent of a coiled-coil region

Optional qualifiers note

* 1. Feature Key COMPBIAS

Definition Extent of a compositionally biased region

Optional qualifiers note

* 1. Feature Key CONFLICT

Definition Different sources report differing sequences

Optional qualifiers note

Comment Examples of values for the “note” qualifier: Missing; K -> Q; GSDSE -> RIRLR; V -> A.

* 1. 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 descriptor in the feature location element is the residue numbers of the cross-linked amino acids in “x..y” format, e.g. “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-methyllanthionine (Cys-Thr);” and “Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in ubiquitin)”

* 1. Feature Key DISULFID

Definition Disulfide bond

Mandatory 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 descriptor in the feature location element is the residue numbers of the linked cysteines in “x..y” format, e.g. “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)”

* 1. Feature Key DNA\_BIND

Definition Extent of a DNA-binding region

Mandatory qualifiers note

Comment The nature of the DNA-binding region is given in the note qualifier. Examples of values for the “note” qualifier: “Homeobox” and “Myb 2”

* 1. Feature Key DOMAIN

Definition Extent of a domain, which is defined as a specific combination of secondary structures organized into a characteristic three-dimensional structure or fold

Mandatory qualifiers note

Comment The domain type is given in the note qualifier. Where several copies of a domain are present, the domains are numbered. Examples of values for the “note” qualifier: “Ras-GAP” and “Cadherin 1”

* 1. Feature Key HELIX

Definition Secondary structure: Helices, for example, Alpha-helix; 3(10) helix; or Pi-helix

Optional qualifiers note

Comment This feature is used only for proteins whose tertiary structure is known. Only three types of secondary structure are specified: helices (key HELIX), beta-strands (key STRAND) and turns (key TURN). Residues not specified in one of these classes are in a ’loop’ or ’random-coil’ structure.

* 1. Feature Key INIT\_MET

Definition Initiator methionine

Optional qualifiers note

Comment The location descriptor in the feature location element is “1”. This feature key indicates the N-terminal methionine is cleaved off. This feature is not used when the initiator methionine is not cleaved off.

* 1. Feature Key INTRAMEM

Definition Extent of a region located in a membrane without crossing it

Optional qualifiers note

* 1. Feature Key LIPID

Definition Covalent binding of a lipid moiety

Mandatory qualifiers note

Comment The chemical nature of the bound lipid moiety is given in the note qualifier, indicating at least the name of the lipidated amino acid. Examples of values for the “note” qualifier: “N-myristoyl glycine”; “GPI-anchor amidated serine” and “S-diacylglycerol cysteine.”

* 1. Feature Key METAL

Definition Binding site for a metal ion.

Mandatory qualifiers note

Comment The note qualifier indicates the nature of the metal. Examples of values for the “note” qualifier: “Iron (heme axial ligand)” and “Copper”.

* 1. Feature Key MOD\_RES

Definition Posttranslational modification of a residue

Mandatory qualifiers note

Comment The chemical nature of the modified residue is given in the note qualifier, indicating at least the name of the post-translationally modified amino acid. If the modified amino acid is listed in Section 4 of this Annex, the abbreviation may be used in place of the the full name. Examples of values for the “note” qualifier: “N-acetylalanine”; “3-Hyp”; and “MeLys” or “N-6-methyllysine"

* 1. Feature Key MOTIF

Definition Short (up to 20 amino acids) sequence motif of biological interest

Optional qualifiers note

* 1. Feature Key MUTAGEN

Definition Site which has been experimentally altered by mutagenesis

Optional qualifiers note

* 1. Feature Key NON\_STD

Definition Non-standard amino acid

Optional qualifiers note

Comment This key only describes the occurrence of non-standard amino acids selenocysteine (U) and pyrrolysine (O) in the amino acid sequence.

* 1. Feature Key NON\_TER

Definition The residue at an extremity of the sequence is not the terminal residue

Optional qualifiers note

Comment If applied to position 1, this means that the first position is not the N-terminus of the complete molecule. If applied to the last position, it means that this position is not the C-terminus of the complete molecule.

* 1. Feature Key NP\_BIND

Definition Extent of a nucleotide phosphate-binding region

Mandatory qualifiers note

Comment The nature of the nucleotide phosphate is indicated in the note qualifier. Examples of values for the “note” qualifier: “ATP” and “FAD”.

* 1. Feature Key PEPTIDE

Definition Extent of a released active peptide

Optional qualifiers note

* 1. Feature Key PROPEP

Definition Extent of a propeptide

Optional qualifiers note

* 1. Feature Key REGION

Definition Extent of a region of interest in the sequence

Optional qualifiers note

* 1. Feature Key REPEAT

Definition Extent of an internal sequence repetition

Optional qualifiers note

* 1. Feature Key SIGNAL

Definition Extent of a signal sequence (prepeptide)

Optional qualifiers note

* 1. 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, or the complete, unabbreviated name of the modified amino acid.

* 1. 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

* 1. Feature Key STRAND

Definition Secondary structure: Beta-strand; for example Hydrogen bonded beta-strand or residue in an isolated beta-bridge

Optional qualifiers note

Comment This feature is used only for proteins whose tertiary structure is known. Only three types of secondary structure are specified: helices (key HELIX), beta-strands (key STRAND) and turns (key TURN). Residues not specified in one of these classes are in a ’loop’ or ’random-coil’ structure.

* 1. Feature Key TOPO\_DOM

Definition Topological domain

Optional qualifiers note

* 1. Feature Key TRANSMEM

Definition Extent of a transmembrane region

Optional qualifiers note

* 1. Feature Key TRANSIT

Definition Extent of a transit peptide (mitochondrion, chloroplast, thylakoid, cyanelle, peroxisome etc.)

Optional qualifiers note

* 1. Feature Key TURN

Definition Secondary structure Turns, for example, H-bonded turn (3-turn, 4-turn or 5-turn)

Optional qualifiers note

Comment This feature is used only for proteins whose tertiary structure is known. Only three types of secondary structure are specified: helices (key HELIX), beta-strands (key STRAND) and turns (key TURN). Residues not specified in one of these classes are in a ’loop’ or ’random-coil’ structure.

* 1. Feature Key UNSURE

Definition Uncertainties in the sequence

Optional qualifiers note

Comment Used to describe region(s) of an amino acid sequence for which the authors are unsure about the sequence presentation.

* 1. Feature Key VARIANT

Definition Authors report that sequence variants exist

Optional qualifiers note

* 1. Feature Key VAR\_SEQ

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

Optional qualifiers note

* 1. 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.

Where the value format is free text that is identified as language dependent, one of the following must be used:

1) the INSDQualifier\_value element; or

2) the NonEnglishQualifier\_value element; or

3) both the INSDQualifier\_value element and the NonEnglishQualifier\_value element.

Where the value format is not identified as language-dependent free text, the INSDQualifier\_value element must be used and the NonEnglishQualifier\_value element must not be used.

PLEASE NOTE: Any qualifier value provided for a qualifier with a language-dependent free text value format may require translation for international, national or regional procedures. The qualifiers listed in the following table are considered to have language-dependent free text values:

Table 6: List of qualifiers for amino acid sequences with language-dependent free text values

|  |  |
| --- | --- |
| **Section** | **Language-Dependent Free Text Qualifier** |
| 8.2 | note |
| 8.3 | organism |

8.1. Qualifier mol\_type

Definition In vivo molecule type of sequence

Mandatory 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

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>Heme (covalent)</INSDQualifier\_value>

Comment The “note” qualifier is mandatory for the feature keys: BINDING; CARBOHYD; CROSSLNK; DISULFID; DNA\_BIND; DOMAIN; LIPID; METAL; MOD\_RES; NP\_BIND; SITE and ZN\_FING

8.3. Qualifier organism

Definition Scientific name of the organism that provided the peptide

Mandatory value format free text  
Language-dependent: this value may require translation for International/National/Regional procedures

Example <INSDQualifier\_value>Homo sapiens</INSDQualifier\_value>

Comment The “organism” qualifier is mandatory for the source feature key.

## SECTION 9: GENETIC CODE TABLES

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

Table 7: Genetic Code Tables

|  |
| --- |
| **1 - Standard Code** |
| AAs = FFLLSSSSYY\*\*CC\*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG  Starts = ---M---------------M---------------M----------------------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **2 - Vertebrate Mitochondrial Code** |
| AAs = FFLLSSSSYY\*\*CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSS\*\*VVVVAAAADDEEGGGG  Starts = --------------------------------MMMM---------------M------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **3 - Yeast Mitochondrial Code** |
| AAs = FFLLSSSSYY\*\*CCWWTTTTPPPPHHQQRRRRIIMMTTTTNNKKSSRRVVVVAAAADDEEGGGG  Starts = ----------------------------------MM---------------M------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **4 - Mold, Protozoan, Coelenterate Mitochondrial Code &**  **Mycoplasma/Spiroplasma Code** |
| AAs = FFLLSSSSYY\*\*CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG  Starts = --MM---------------M------------MMMM---------------M------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **5 - Invertebrate Mitochondrial Code** |
| AAs = FFLLSSSSYY\*\*CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSSSSVVVVAAAADDEEGGGG  Starts = ---M----------------------------MMMM---------------M------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **6 - Ciliate, Dasycladacean and Hexamita Nuclear Code** |
| AAs = FFLLSSSSYYQQCC\*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG  Starts = -----------------------------------M----------------------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |

|  |
| --- |
| **9 - Echinoderm and Flatworm Mitochondrial Code** |
| AAs = FFLLSSSSYY\*\*CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNNKSSSSVVVVAAAADDEEGGGG  Starts = -----------------------------------M---------------M------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **10 - Euplotid Nuclear Code** |
| AAs = FFLLSSSSYY\*\*CCCWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG  Starts = -----------------------------------M----------------------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **11 – Bacterial, Archaeal, and Plant Plastid Code** |
| AAs = FFLLSSSSYY\*\*CC\*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG  Starts = ---M---------------M------------MMMM---------------M------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **12 - Alternative Yeast Nuclear Code** |
| AAs = FFLLSSSSYY\*\*CC\*WLLLSPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG  Starts = -------------------M---------------M----------------------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **13 - Ascidian Mitochondrial Code** |
| AAs = FFLLSSSSYY\*\*CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSSGGVVVVAAAADDEEGGGG  Starts = ---M------------------------------MM---------------M------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **14 - Alternative Flatworm Mitochondrial Code** |
| AAs = FFLLSSSSYYY\*CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNNKSSSSVVVVAAAADDEEGGGG  Starts = -----------------------------------M----------------------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **16 - Chlorophycean Mitochondrial Code** |
| AAs = FFLLSSSSYY\*LCC\*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG  Starts = -----------------------------------M----------------------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |

|  |
| --- |
| **21 - Trematode Mitochondrial Code** |
| AAs = FFLLSSSSYY\*\*CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNNKSSSSVVVVAAAADDEEGGGG  Starts = -----------------------------------M---------------M------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **22 - Scenedesmus obliquus Mitochondrial Code** |
| AAs = FFLLSS\*SYY\*LCC\*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG  Starts = -----------------------------------M----------------------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **23 - Thraustochytrium Mitochondrial Code** |
| AAs = FF\*LSSSSYY\*\*CC\*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG  Starts = --------------------------------M--M---------------M------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **24 - Pterobranchia Mitochondrial Code** |
| AAs = FFLLSSSSYY\*\*CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSSKVVVVAAAADDEEGGGG  Starts = ---M---------------M---------------M---------------M------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **25 - Candidate Division SR1 and Gracilibacteria Code** |
| AAs = FFLLSSSSYY\*\*CCGWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG  Starts = ---M-------------------------------M---------------M------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **26 - Pachysolen tannophilus Nuclear Code** |
| AAs = FFLLSSSSYY\*\*CC\*WLLLAPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG  Starts = -------------------M---------------M----------------------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **27 - Karyorelict Nuclear Code** |
| AAs = FFLLSSSSYYQQCCWWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG  Starts = --------------\*--------------------M----------------------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |

|  |
| --- |
| **28 - Condylostoma Nuclear Code** |
| AAs = FFLLSSSSYYQQCCWWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG  Starts = ----------\*\*--\*--------------------M----------------------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **29 - Mesodinium Nuclear Code** |
| AAs = FFLLSSSSYYYYCC\*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG  Starts = -----------------------------------M----------------------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **30 - Peritrich Nuclear Code** |
| AAs = FFLLSSSSYYEECC\*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG  Starts = -----------------------------------M----------------------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **31 - Blastocrithidia Nuclear Code** |
| AAs = FFLLSSSSYYEECCWWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG  Starts = ----------\*\*-----------------------M----------------------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **33 - Cephalodiscidae Mitochondrial UAA-Tyr Code** |
| AAs = FFLLSSSSYYY\*CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSSKVVVVAAAADDEEGGGG  Starts = ---M---------------M---------------M---------------M------------  Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg  Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg  Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |

[Annex II follows]

# **ANNEX II**

DOCUMENT TYPE DEFINITION (DTD) FOR SEQUENCE LISTING

*Version 1.3*

*Proposal presented for consideration by the Committee on WIPO Standards (CWS)*

*at its tenth session*

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

<!--Annex II of WIPO Standard ST.26, Document Type Definition (DTD) for Sequence Listing

This entity may be identified by the PUBLIC identifier:

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

PUBLIC "-//WIPO//DTD SEQUENCE LISTING 1.3//EN" "ST26SequenceListing\_V1\_3.dtd"

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* PUBLIC DTD URL

\* https://www.wipo.int/standards/dtd/ST26SequenceListing\_V1\_3.dtd

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Revision of Annex II to WIPO Standard ST.26 was approved by the Committee on WIPO

\* Standards (CWS) at its eighth session.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* CONTACTS

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*

\* xml.standards@wipo.int

\*

\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* NOTES

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* The sequence data part is a subset of the complete INSDC DTD V.1.5 that only covers

\* the requirements of WIPO Standard ST.26.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* REVISION HISTORY

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

2021-11-05: Revised Version 1.3 approved at CWS/9 (small edits to the comments)

2020-05-20: Version 1.3 approved at CWS/8.

Changes:

- Optional originalFreeTextLanguageCode attribute added to <ST26SequenceListing> to allow applicants to indicate the language of the free text in the original sequence listing.

- Optional nonEnglishFreeTextLanguageCode attribute added to <ST26SequenceListing> to allow applicants to indicate the language of the free text provided in the element <NonEnglishQualifier\_value>.

- Optional id attribute added to INSDQualifier to facilitate comparison of language-dependent qualifier values between sequence listings.

- Optional element <NonEnglishQualifier\_value> added to element <INSDQualifier> to allow applicants to type language-dependent qualifiers in a non-English Language with the characters set forth in paragraph 40(a) of the ST.26 main body document.

2018-10-19: Version 1.2 approved at CWS/6.

Changes:

<INSDQualifier\*> changed to <INSDQualifier+> for alignment with business needs and advice from NCBI (an INSDFeature\_quals element (if present) should have one or more INSDQualifier elements).

2017-06-02: Version 1.1 approved at the CWS/5

Changes:

Comments added to <INSDSeq\_length>, <INSDSeq\_division> and <INSDSeq\_sequence> to clarify the reason of the differences between the INSDC DTD v.1.5 and ST26 Sequence Listing DTD V1\_1.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

2016-03-24: Version 1.0 adopted at the CWS/4Bis

2014-03-11: Final draft for adoption.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

ST26SequenceListing

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* ROOT ELEMENT

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

-->

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

<!--The elements ApplicantName and InventorName are optional in this DTD to facilitate

the conversion between various encoding schemes-->

<!--originalFreeTextLanguageCode:

The language code (see reference in paragraph 9 to ISO 639-1:2002) for the single original language in which the language-dependent free text qualifiers (NonEnglishQualifier\_value) were prepared.

-->

<!--nonEnglishFreeTextLanguageCode:

The language code (see reference in paragraph 9 to ISO 639-1:2002) for the language in which the language-dependent free text qualifiers (NonEnglishQualifier\_value) currently correspond.

-->

<!--fileName:

By default this will be set

to the value provided for the project name in WIPO Sequence. If the value is identical to

the actual ST.26 XML filename, it should be noted that Offices may enforce their

requirements for the filename used which may restrict which characters are allowable for

submitted electronic files. It is also acceptable for the value of the filename attribute and the actual file name to be different. Please refer to the WIPO Sequence and ST.26 Knowledge Base for further details on Offices’ naming conventions for electronic files

--->

<!ATTLIST ST26SequenceListing

dtdVersion CDATA #REQUIRED

fileName CDATA #IMPLIED

softwareName CDATA #IMPLIED

softwareVersion CDATA #IMPLIED

productionDate CDATA #IMPLIED

originalFreeTextLanguageCode CDATA #IMPLIED

nonEnglishFreeTextLanguageCode CDATA #IMPLIED

>

<!--ApplicantFileReference

Applicant's or agent's file reference, mandatory if application identification not provided.

-->

<!ELEMENT ApplicantFileReference (#PCDATA)>

<!--ApplicationIdentification

Application identification for which the sequence listing is submitted, when available.

-->

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

<!--EarliestPriorityApplicationIdentification

Identification of the earliest priority application, which contains IPOfficeCode, ApplicationNumberText and FilingDate elements.

-->

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

<!--ApplicantName

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

-->

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

-->

<!ELEMENT ApplicantName (#PCDATA)>

<!ATTLIST ApplicantName

languageCode CDATA #REQUIRED

>

<!--ApplicantNameLatin

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

-->

<!ELEMENT ApplicantNameLatin (#PCDATA)>

<!--InventorName

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

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

-->

<!ELEMENT InventorName (#PCDATA)>

<!ATTLIST InventorName

languageCode CDATA #REQUIRED

>

<!--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).

-->

<!ELEMENT InventorNameLatin (#PCDATA)>

<!--InventionTitle

Title of the invention typed in the characters as set forth in paragraph 40(a) in the language of filing. A translation of the title of the invention into additional languages may be typed in the characters as set forth in paragraph 40(a) using additional InventionTitle elements. The title of invention should be between two to seven words.

-->

<!--languageCode: Appropriate language code from ISO 639-1 - Codes

for the representation of names of languages - Part 1: Alpha-2

-->

<!ELEMENT InventionTitle (#PCDATA)>

<!ATTLIST InventionTitle

languageCode CDATA #REQUIRED

>

<!--SequenceTotalQuantity

Indicates the total number of sequences in the document.

Its purpose is to be quickly accessible for automatic processing.

-->

<!ELEMENT SequenceTotalQuantity (#PCDATA)>

<!--SequenceData

Data for individual Sequence.

For intentionally skipped sequences see the ST.26 main body document.

-->

<!ELEMENT SequenceData (INSDSeq)>

<!ATTLIST SequenceData

sequenceIDNumber CDATA #REQUIRED

>

<!--IPOfficeCode

ST.3 code. For example, if the application identification is PCT/IB2013/099999, then IPOfficeCode value will be "IB" for the International Bureau of WIPO.

-->

<!ELEMENT IPOfficeCode (#PCDATA)>

<!--ApplicationNumberText

The application identification as provided by the office of filing (e.g. PCT/IB2013/099999)

-->

<!ELEMENT ApplicationNumberText (#PCDATA)>

<!--FilingDate

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

-->

<!ELEMENT FilingDate (#PCDATA)>

<!--\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* INSD Part

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

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

The INSD part is subset of the INSD DTD v1.5 and as such can only be used to generate an XML instance as it will not support the complete INSD structure.

This part is based on:

The International Nucleotide Sequence Database (INSD) collaboration.

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

-->

<!--INSDSeq

Sequence data. Changed INSD V1.5 DTD elements, INSDSeq\_division and INSDSeq\_sequence from optional to mandatory per business requirements.

-->

<!ELEMENT INSDSeq (INSDSeq\_length, INSDSeq\_moltype, INSDSeq\_division, INSDSeq\_other-seqids?, INSDSeq\_feature-table?, INSDSeq\_sequence)>

<!--INSDSeq\_length

The length of the sequence. INSDSeq\_length allows only integer.

-->

<!ELEMENT INSDSeq\_length (#PCDATA)>

<!--INSDSeq\_moltype

Admissible values: DNA, RNA, AA

-->

<!ELEMENT INSDSeq\_moltype (#PCDATA)>

<!--INSDSeq\_division

Indication that a sequence is related to a patent application. Must be populated with the value PAT.

-->

<!ELEMENT INSDSeq\_division (#PCDATA)>

<!--INSDSeq\_other-seqids

In the context of data exchange with database providers, the IPOs should populate for each sequence the element INSDSeq\_other-seqids with one INSDSeqid containing a reference to the corresponding published patent and the sequence identification.

-->

<!ELEMENT INSDSeq\_other-seqids (INSDSeqid?)>

<!--INSDSeq\_feature-table

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

-->

<!ELEMENT INSDSeq\_feature-table (INSDFeature+)>

<!--INSDSeq\_sequence

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

-->

<!ELEMENT INSDSeq\_sequence (#PCDATA)>

<!--INSDSeqid

Intended for the use of IPOs 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)

-->

<!ELEMENT INSDSeqid (#PCDATA)>

<!--INSDFeature

Description of one feature.

-->

<!ELEMENT INSDFeature (INSDFeature\_key, INSDFeature\_location, INSDFeature\_quals?)>

<!--INSDFeature\_key

A word or abbreviation indicating a feature.

-->

<!ELEMENT INSDFeature\_key (#PCDATA)>

<!--INSDFeature\_location

Region of the presented sequence which corresponds to the feature.

-->

<!ELEMENT INSDFeature\_location (#PCDATA)>

<!--INSDFeature\_quals

List of qualifiers containing auxiliary information about a feature.

-->

<!ELEMENT INSDFeature\_quals (INSDQualifier+)>

<!--INSDQualifier

Additional information about a feature.

For coding sequences and variants see the ST.26 main body document.

-->

<!--id

Unique identifier for the INSDQualifier to facilitate comparison of versions of a sequence listing specifically having language-dependent qualifier values in different languages.

-->

<!ELEMENT INSDQualifier (INSDQualifier\_name, INSDQualifier\_value?, NonEnglishQualifier\_value?)>

<!ATTLIST INSDQualifier

id ID #IMPLIED

>

<!--INSDQualifier\_name

Name of the qualifier.

-->

<!ELEMENT INSDQualifier\_name (#PCDATA)>

<!--INSDQualifier\_value

Value of the qualifier. Where the qualifier is language-dependent its value must be in the English language and typed with the characters set forth in paragraph 40 (b).

-->

<!ELEMENT INSDQualifier\_value (#PCDATA)>

<!--NonEnglishQualifier\_value

Value of a language-dependent qualifier in a language that is not English and typed with the characters set forth in paragraph 40 (a). The language is indicated with the attribute nonEnglishFreeTextLanguageCode.

-->

<!ELEMENT NonEnglishQualifier\_value (#PCDATA)>

[Annex III follows]

# **ANNEX III**

SEQUENCE LISTING SPECIMEN (XML file)

*Version 1.4*

*Proposal presented for consideration by the Committee on WIPO Standards (CWS)*

*at its tenth session*

The Annex III is available at: <https://www.wipo.int/standards/en/xml_material/st26/st26-annex-iii-sequence-listing-specimen.xml>

[Annex IV follows]

# **ANNEX IV**

CHARACTER SUBSET FROM THE UNICODE BASIC LATIN CODE TABLE  
FOR USE IN AN XML INSTANCE OF A SEQUENCE LISTING

*Version 1.3*

*Proposal presented for consideration by the Committee on WIPO Standards (CWS)  
at its tenth session*

The ampersand character (0026) is only permitted as part of a predefined entity. The quotation mark (0022), the apostrophe (0027), the less-than sign (003C), and the greater-than sign (003E) must be represented by their predefined entities. In addition, the ampersand character (0026) must be represented by its predefined entity when used as an ampersand in a value of an attribute or content of an element.

| **Unicode code point** | **Character** | **Name** |
| --- | --- | --- |
| 0020 |  | SPACE |
| 0021 | ! | EXCLAMATION MARK |
| 0022 | “ | QUOTATION MARK |
| 0023 | # | NUMBER SIGN |
| 0024 | $ | DOLLAR SIGN |
| 0025 | % | PERCENT SIGN |
| 0026 | & | AMPERSAND |
| 0027 | ‘ | APOSTROPHE |
| 0028 | ( | LEFT PARENTHESIS |
| 0029 | ) | RIGHT PARENTHESIS |
| 002A | \* | ASTERISK |
| 002B | + | PLUS SIGN |
| 002C | , | COMMA |
| 002D | - | HYPHEN-MINUS |
| 002E | . | FULL STOP |
| 002F | / | SOLIDUS |
| 0030 | 0 | DIGIT ZERO |
| 0031 | 1 | DIGIT ONE |
| 0032 | 2 | DIGIT TWO |
| 0033 | 3 | DIGIT THREE |
| 0034 | 4 | DIGIT FOUR |
| 0035 | 5 | DIGIT FIVE |
| 0036 | 6 | DIGIT SIX |
| 0037 | 7 | DIGIT SEVEN |
| 0038 | 8 | DIGIT EIGHT |
| 0039 | 9 | DIGIT NINE |
| 003A | : | COLON |
| 003B | ; | SEMICOLON |
| 003C | < | LESS-THAN-SIGN |
| 003D | = | EQUALS SIGN |
| 003E | > | GREATER-THAN-SIGN |
| 003F | ? | QUESTION MARK |
| 0040 | @ | COMMERCIAL AT |
| 0041 | A | LATIN CAPITAL LETTER A |
| 0042 | B | LATIN CAPITAL LETTER B |
| 0043 | C | LATIN CAPITAL LETTER C |
| 0044 | D | LATIN CAPITAL LETTER D |
| 0045 | E | LATIN CAPITAL LETTER E |
| 0046 | F | LATIN CAPITAL LETTER F |
| 0047 | G | LATIN CAPITAL LETTER G |
| 0048 | H | LATIN CAPITAL LETTER H |
| 0049 | I | LATIN CAPITAL LETTER I |
| 004A | J | LATIN CAPITAL LETTER J |
| 004B | K | LATIN CAPITAL LETTER K |
| 004C | L | LATIN CAPITAL LETTER L |
| 004D | M | LATIN CAPITAL LETTER M |
| 004E | N | LATIN CAPITAL LETTER N |
| 004F | O | LATIN CAPITAL LETTER O |
| 0050 | P | LATIN CAPITAL LETTER P |
| 0051 | Q | LATIN CAPITAL LETTER Q |
| 0052 | R | LATIN CAPITAL LETTER R |
| 0053 | S | LATIN CAPITAL LETTER S |
| 0054 | T | LATIN CAPITAL LETTER T |
| 0055 | U | LATIN CAPITAL LETTER U |
| 0056 | V | LATIN CAPITAL LETTER V |
| 0057 | W | LATIN CAPITAL LETTER W |
| 0058 | X | LATIN CAPITAL LETTER X |
| 0059 | Y | LATIN CAPITAL LETTER Y |
| 005A | Z | LATIN CAPITAL LETTER Z |
| 005B | [ | LEFT SQUARE BRACKET |
| 005C | \ | REVERSE SOLIDUS |
| 005D | ] | RIGHT SQUARE BRACKET |
| 005E | ^ | CIRCUMFLEX ACCENT |
| 005F | \_ | LOW LINE |
| 0060 | ` | GRAVE ACCENT |
| 0061 | a | LATIN SMALL LETTER A |
| 0062 | b | LATIN SMALL LETTER B |
| 0063 | c | LATIN SMALL LETTER C |
| 0064 | d | LATIN SMALL LETTER D |
| 0065 | e | LATIN SMALL LETTER E |
| 0066 | f | LATIN SMALL LETTER F |
| 0067 | g | LATIN SMALL LETTER G |
| 0068 | h | LATIN SMALL LETTER H |
| 0069 | i | LATIN SMALL LETTER I |
| 006A | j | LATIN SMALL LETTER J |
| 006B | k | LATIN SMALL LETTER K |
| 006C | l | LATIN SMALL LETTER L |
| 006D | m | LATIN SMALL LETTER M |
| 006E | n | LATIN SMALL LETTER N |
| 006F | o | LATIN SMALL LETTER O |
| 0070 | p | LATIN SMALL LETTER P |
| 0071 | q | LATIN SMALL LETTER Q |
| 0072 | r | LATIN SMALL LETTER R |
| 0073 | s | LATIN SMALL LETTER S |
| 0074 | t | LATIN SMALL LETTER T |
| 0075 | u | LATIN SMALL LETTER U |
| 0076 | v | LATIN SMALL LETTER V |
| 0077 | w | LATIN SMALL LETTER W |
| 0078 | x | LATIN SMALL LETTER X |
| 0079 | y | LATIN SMALL LETTER Y |
| 007A | z | LATIN SMALL LETTER Z |
| 007B | { | LEFT CURLY BRACKET |
| 007C | | | VERTICAL LINE |
| 007D | } | RIGHT CURLY BRACKET |
| 007E | ~ | TILDE |

[Annex V follows]

# **ANNEX V**

ADDITIONAL DATA EXCHANGE REQUIREMENTS (FOR IPOs ONLY)

*Version 1.4*

*Proposal presented for consideration by the Committee on WIPO Standards (CWS)*

*at its tenth session*

In the context of data exchange with database providers (INSD members), the IPOs 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).

[Annex VI follows]

# **ANNEX VI**

GUIDANCE DOCUMENT WITH ILLUSTRATED EXAMPLES

*Version 1.6*

*Proposal presented for consideration by the Committee on WIPO Standards (CWS)*

*at its tenth session*

TABLE OF CONTENTS

INTRODUCTION………….……………………………………………………………………….…………………….[3.26.vi.1](#AnnexVIIntroduction)

EXAMPLE INDEX……………………………………………………………………………………….……………….[3.26.vi.8](#AnnexVIExampleIndex)

EXAMPLES………………………………………………..………………………………………….………………….[3.26.vi.18](#AnnexVIExamples)

APPENDIX………………………………………….....………………………………………………………………….[3.26.vi.76](#AnnexVIAppendix)

## INTRODUCTION

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

This guidance document consists of this introduction, an example index, examples of sequence disclosures, and an appendix containing a sequence listing in XML with sequences from the examples. This introduction explains certain concepts and terminology used in the remainder of this document. The examples illustrate the requirements of specific paragraphs of the Standard and each example has been designated with the most relevant paragraph number. Some examples further illustrate other paragraphs and appropriate cross-references are indicated at the end of each example. The index provides page numbers for the examples and any indicated cross-references. Each sequence in an example that either must or may be included in a sequence listing has been assigned a sequence identification number (SEQ ID NO) and appears in XML format in the [Appendix](#Appendix) to this document.

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

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

### *Preparation of a sequence listing*

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

1. Does ST.26 paragraph 7 require inclusion of a particular disclosed sequence?

2. If inclusion of a particular disclosed sequence is not required, is inclusion of that sequence permitted by ST.26?

3. If inclusion of a particular disclosed sequence is required or permitted by ST.26, how should that sequence be represented in the sequence listing?

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

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

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

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

#### Enumeration of its residues

ST.26 paragraph 3(c) defines “enumeration of its residues” as disclosure of a sequence in a patent application by listing, in order, each residue of the sequence, wherein (i) the residue is represented by a name, abbreviation, symbol, or structure; or (ii) multiple residues are represented by a shorthand formula. A sequence should be disclosed in a patent application by “enumeration of its residues” using conventional symbols, which are the nucleotide symbols set forth in Section 1, Table 1 of ST.26 Annex 1 (i.e., the lower case symbols or their upper case equivalents[[2]](#footnote-3)) and the amino acid symbols set forth in Section 3, Table 3 of ST.26 Annex 1 (i.e., the upper case symbols or their lower case equivalents1). Hereinafter, these nucleotide and amino acid symbols are referred to as conventional symbols. Representations of nucleotides and amino acids that are other than those set forth in these tables are hereinafter referred to as “nonconventional”.

Where a representation of a residue is disclosed as equivalent to a conventional symbol or abbreviation (e.g., “Z1” means “A”), or to a specific sequence of conventional symbols (e.g., “Z1” means “agga”), then the sequence is interpreted as though it were disclosed using the equivalent conventional symbol(s) or abbreviation(s), to determine whether ST.26 paragraph 7 requires inclusion in the sequence listing or whether paragraph 8 prohibits inclusion. Where a nonconventional nucleotide symbol is used as an ambiguity symbol (e.g., X1 = inosine or pseudouridine), but is not equivalent to one of the conventional ambiguity symbols in Section 1, Table 1 (i.e., “m”, “r”, “w”, “s”, “y”, “k”, “v”, “h”, “d”, “b”, or “n”), then the residue is interpreted as an “n” residue to determine whether ST.26 Paragraph 7 requires inclusion of the sequence in the sequence listing or whether ST.26 Paragraph 8 prohibits inclusion. Similarly, where a nonconventional amino acid symbol is used as an ambiguity symbol (e.g., “Z1” means “A”, “G”, “S” or “T”), but is not equivalent to one of the conventional ambiguity symbols in Section 3, Table 3 (i.e., B, Z, J, or X), then the residue is interpreted as an “X” residue to determine whether ST.26 paragraph 7 requires inclusion of the sequence in the sequence listing or whether ST.26 paragraph 8 prohibits inclusion.

Care should be taken to disclose sequences using conventional symbols; however, where sequences are otherwise disclosed, it may be necessary to consult the disclosure for an explanation to determine the meaning of the nonconventional representation.

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

#### Specifically defined

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

#### Most encompassing sequence

Where a sequence that meets the requirements of paragraph 7 is disclosed by enumeration of its residues only once in an application, but is described differently in multiple embodiments, e.g., one embodiment “X” in one or more locations could be any amino acid, but in further embodiments, “X” could be only a limited number of amino acids, ST.26 requires inclusion in a sequence listing of only the single sequence that has been enumerated by its residues. As per paragraphs 15 and 27, where such a sequence contains multiple “n” or “X” ambiguity symbols, “n” or “X” is construed to represent any nucleotide or amino acid, respectively, in the absence of further annotation. Consequently, the single sequence required to be included is the most encompassing sequence disclosed. The most encompassing sequence is the single sequence having variant residues that are represented by the most restrictive ambiguity symbols that include the most disclosed embodiments. Likewise, where a sequence is disclosed by enumeration of its residues only once, but the length of the sequence may vary due to copy number variation, the longest embodiment of the sequence is considered the most encompassing sequence. For example, consider a sequence containing a repeated region that can vary from 2 to 5 copies as enumerated. The embodiment with 5 copies of the repeat is the most encompassing sequence and should be included in the sequence listing. However, inclusion of additional specific sequences is strongly encouraged where practical, e.g., those that represent additional embodiments that are a key part of the invention. Inclusion of the additional sequences allows for a more thorough search and provides public notice of the subject matter for which a patent is sought.

### *Usage of Ambiguity Symbol*

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

The symbol “n”

* 1. must not be used to represent anything other than a single nucleotide;
  2. will be construed as any one of “a”, “c”, “g”, or “t/u” except where it is used with a further description;
  3. should be used to represent any of the following nucleotides together with a further description:
     1. modified nucleotide, e.g., natural, synthetic, or non-naturally occurring, that cannot otherwise be represented by any other symbol in Annex I (see Section 1, Table 1);
     2. “unknown” nucleotide, i.e., not determined, not disclosed, or unsure;
     3. an abasic site; or
  4. may be used to represent a sequence variant, i.e., alternatives, deletions, insertions, or substitutions, where “n” is the most restrictive ambiguity symbol.

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

The symbol “X”

1. must not be used to represent anything other than a single amino acid;
2. 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;
3. should be used to represent any of the following amino acids together with a further description:
   * 1. modified amino acid, e.g., natural, synthetic, or non-naturally occurring, that cannot otherwise be represented by any other symbol in Annex I (see Section 3, Table 3);
     2. “unknown” amino acid, i.e., not determined, not disclosed, or unsure; or
4. may be used to represent a sequence variant, i.e., alternatives, deletions, insertions, or substitutions, where “X” is the most restrictive ambiguity symbol.

#### Annotation of Modified Residues

This Standard requires that “modified” residues are annotated per paragraph 17 for nucleotides, and per paragraph 30 for amino acids.

ST.26 paragraph 3(e) defines “modified amino acid” as 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. Similarly, the Standard defines “modified nucleotide” as any nucleotide as described in paragraph 3(g) 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 (ST.26, paragraph 3(f)).

Based on the definitions above, modifications to the nucleobases or sugar-phosphate backbone of a nucleic acid and modifications to the amnio acid R groups or peptide backbone of a peptide result in one or more “modified nucleotides” or “modified amino acids,” respectively. Therefore, such nucleotides and amino acids must be annotated. Examples of backbone modifications include nucleotide analogs such as peptide nucleic acids (PNAs) and glycol nucleic acids (GNAs), and D-amino acids.

Note that modification of a terminal amino acid of a peptide or a terminal nucleotide of a nucleic acid does not necessarily result in a “modified amino acid” or “modified nucleotide”. One must look at the terminal modification and determine whether the modification changes the chemical structure of residue such that residue falls outside the exceptions set forth within paragraph 3(e) and 3(f). For example, a peptide in which the C terminal residue is linked to a structure (such as part of a branched sequence – see peptide #2 in example 7(b)-3) via a conventional amide linkage is not considered a “modified residue” and therefore is not required to be annotated. Similarly, a peptide in which the N terminal residue is amide bonded to biotin is not considered a “modified residue” and therefore is not required to be annotated. In both scenarios, the structure of the residue involved in the C-terminal or N-terminal linkage is not changed from the conventional amino acids recited in paragraph 3(e) of the Standard.

In contrast, terminal modifications that change the chemical structure of the residue are considered “modified residues” and must be annotated. For example, the methylation of the C-terminus in Example 3(c)-1 does change the chemical structure of the terminal residue, since the methyl group replaces the hydroxyl normally found at the alpha carboxyl group. Therefore, this methylated lysine must be annotated as a “modified residue”.

Note that it will be up to the applicant to evaluate each terminal residue modification within an enumerated sequence and make a determination as to whether or not the structure of the terminal residue is changed. If the modified residue structure is different from the conventional amino acids or nucleotides indicated in paragraph 3(e) and 3(f) of the Standard, then the modification must be annotated.

Finally, it is always recommended that applicants include as much information as reasonable in their sequence listings to represent their disclosures as accurately as possible. Therefore, even if a modification isn’t required to be annotated, it should preferably be included.

Note however that annotation of variants of an enumerated, primary sequence must comply with the requirements of in ST.26 paragraphs 93-100. Modifications that are disclosed as variants of an enumerated sequence may not be required to be included in the sequence listing. For the definition of annotation of variants, see ST.26 paragraphs 93-95.

#### Representation of Modified Residues

ST.26 indicates that modified nucleotides and amino acids should be represented in the sequence listing as the corresponding unmodified residue whenever possible (see paragraphs 16 and 29). Note that this recommendation is a “should” – a “strongly encouraged approach, but not a requirement” (see paragraph 4(d)). It is up to the discretion of the applicant to decide if a modified residue will be represented by the corresponding unmodified residue or the variables “n” or “X”.

As a general rule of thumb – if a residue is modified by the addition of a moiety, such as methylation or acetylation, and the structure of the unmodified residue is generally unchanged, then representation by the unmodified residue is recommended. For example, a methylated adenosine should preferably be represented by “a” in the sequence listing. However, when the modified residue is structurally different from any unmodified residue, then an “n” or an “X” is recommended. For example, norleucine is an isomer of leucine, and its side chain is a linear structure of 4 carbons. Leucine also has a 4 carbon side chain, but it is branched at the second carbon. Therefore, norleucine isn’t simply the result of a modification added to a leucine, but a completely different (although related) structure. It is therefore recommended that Norleucine be represented by an “X” in a sequence listing.

A nucleotide is “specifically defined” when it is represented by anything other than ‘n’, and an amino acid is “specifically defined” when it is represented by anything other than ‘X’ (see ST.26, paragraph 3(k)). Therefore, a 2’ O-methyl adenosine represented by an ‘a’ in the sequence is specifically defined, whereas norleucine represented by ‘X’ in the sequence is not specifically defined.

|  |  |
| --- | --- |
| **Symbol** | **Definition** |
| a | adenine |
| c | cytosine |
| g | guanine |
| t | thymine in DNA  uracil in RNA (t/u) |
| m | a or c |
| r | a or g |
| w | a or t/u |
| s | c or g |
| y | c or t/u |
| k | g or t/u |
| v | a or c or g; not t/u |
| h | a or c or t/u; not g |
| d | a or g or t/u; not c |
| b | c or g or t/u; not a |
| n | a or c or g or t/u; “unknown” or “other” |

### *Table A – Conventional Nucleotide Symbols, and Definitions*

|  |  |  |
| --- | --- | --- |
| **Symbol** | **3-Letter Code** | **Definition** |
| A | Ala | Alanine |
| R | Arg | Arginine |
| N | Asn | Asparagine |
| D | Asp | Aspartic Acid (Aspartate) |
| C | Cys | Cysteine |
| Q | Gln | Glutamine |
| E | Glu | Glutamic Acid (Glutamate) |
| G | Gly | Glycine |
| H | His | Histidine |
| I | Ile | Isoleucine |
| L | Leu | Leucine |
| K | Lys | Lysine |
| M | Met | Methionine |
| F | Phe | Phenylalanine |
| P | Pro | Proline |
| O | Pyl | Pyrrolysine |
| S | Ser | Serine |
| U | Sec | Selenocysteine |
| T | Thr | Threonine |
| W | Trp | Tryptophan |
| Y | Tyr | Tyrosine |
| V | Val | Valine |
| B | Asx | Aspartic Acid or Asparagine |
| Z | Glx | Glutamine or Glutamic Acid |
| J | Xle | Leucine or Isoleucine |
| X | Xaa | A or R or N or D or C or Q or E or G or H or I or L or K or M or F or P or O or S or U or T or W or Y or V, “unknown” or “other” |

### *Table B – Conventional Amino Acid Symbols, Three letter Codes, and Definitions*

EXAMPLE INDEX

**Page**

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

**Example 3(a)-1: D-amino acids** [18](#page18_3a1)

***Cross-referenced examples***

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid [56](#page56_29_1)

Example 29-2: Use of the corresponding unmodified amino acid………………………………….....………….[57](#page57_29_2)

Example 30-1: Feature key “CARBOHYD” [58](#page58_30_1)

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

**Example 3(c)-1: Enumeration of amino acids by chemical structure** [**19**](#page19_3c1)

**Example 3(c)-2: Shorthand formula for an amino acid sequence** [**20**](#page20_3c2)

***Cross-referenced examples***

Example 27-1: Shorthand formula for an amino acid sequence [51](#page51_27_1)

Example 27-3: Shorthand formula - four or more specifically defined amino acids [53](#page53_27_3)

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

***Cross-referenced examples***

Example 3(g)-4: Nucleic Acid Analogues [24](#page24_3g4)

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

**Example 3(g)-1: Nucleotide sequence interrupted by a C3 spacer** [**21**](#page21_3g1)

**Example 3(g)-2: Nucleotide sequence with residue alternatives, including a C3 spacer** [**22**](#page22_3g2)

**Example 3(g)-3: Abasic site** [**23**](#page23_3g3)

**Example 3(g)-4: Nucleic Acid Analogues** [**24**](#page24_3g4)

***Cross-referenced examples***

Example 11(b)-1: Double-stranded nucleotide sequence – different lengths [47](#page47_11b1)

Example 14-1: The symbol “t” represents uracil in RNA [49](#page49_14_1)

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

**Example 3(k)-1: Nucleotide ambiguity symbols** [**25**](#page25_3k1)

**Example 3(k)-2: Ambiguity symbol “n” used in both a conventional and nonconventional manner** [**26**](#page26_3k2)

**Example 3(k)-3: Ambiguity symbol “n” used in a nonconventional manner** [**27**](#page27_3k3)

**Example 3(k)-4: Ambiguity symbols other than “n” are “specifically defined”** [**28**](#page28_3k4)

**Example 3(k)-5: Ambiguity abbreviation “Xaa” used in a nonconventional manner** [**29**](#page29_3k5)

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

***Cross-referenced examples***

Example 28-1: Encoding nucleotide sequence and encoded amino acid sequence [54](#page54_28_1)

Example 55-1: Combined DNA/RNA Molecule [64](#page64_55_1)

Example 89-2: Feature location extends beyond the disclosed sequence [66](#page66_89_2)

Example 92-1: Amino acid sequence encoded by a coding sequence with introns [68](#page68_92_1)

**Paragraph 7(a) – Nucleotide sequences required in a sequence listing**

**Example 7(a)-1: Branched nucleotide sequence** [**30**](#page30_7a1)

**Example 7(a)-2: Linear nucleotide sequence having a secondary structure** [**32**](#page32_7a2)

**Example 7(a)-3: Nucleotide ambiguity symbols used in a nonconventional manner** [**33**](#page33_7a3)

**Example 7(a)-4: Nucleotide ambiguity symbols used in a nonconventional manner** [**34**](#page34_7a4)

**Example 7(a)-5: Nonconventional nucleotide symbols** [**35**](#page35_7a5)

**Example 7(a)-6: Nonconventional nucleotide symbols** [**36**](#page36_7a6)

***Cross-referenced examples***

Example 3(g)-1: Nucleotide sequence interrupted by a C3 spacer [21](#page21_3g1)

Example 3(g)-2: Nucleotide sequence with residue alternatives, including a C3 spacer [22](#page22_3g2)

Example 3(g)-3: Abasic site [23](#page23_3g3)

Example 3(g)-4: Nucleic Acid Analogues [24](#page24_3g4)

Example 3(k)-1: Nucleotide ambiguity symbols [25](#page25_3k1)

Example 3(k)-2: Ambiguity symbol “n” used in both a conventional and nonconventional manner [26](#page26_3k2)

Example 3(k)-3: Ambiguity symbol “n” used in a nonconventional manner [27](#page27_3k3)

Example 3(k)-4: Ambiguity symbols other than “n” are “specifically defined” [28](#page28_3k4)

Example 11(a)-1: Double-stranded nucleotide sequence – same lengths [46](#page46_11a1)

Example 11(b)-1: Double-stranded nucleotide sequence – different lengths [47](#page47_11b1)

Example 11(b)-2: Double-stranded nucleotide sequence – no base-pairing segment [48](#page48_11b2)

Example 14-1: The symbol “t” represents uracil in RNA [49](#page49_14_1)

Example 89-1: Encoding nucleotide sequence and encoded amino acid sequence [65](#page65_89_1)

Example 93-1: Representation of enumerated variants [70](#page70_93_1)

Example 95(b)-1: Representation of individual variant sequences

with multiple interdependent variations [75](#page75_95b1)

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

**Example 7(b)-1: Four or more specifically defined amino acids** [**37**](#page37_7b1)

**Example 7(b)-2: Branched amino acid sequence** [**38**](#page38_7b2)

**Example 7(b)-3: Branched amino acid sequence** [**41**](#page41_7b3)

**Example 7(b)-4: Cyclic peptide containing a branched amino acid sequence** [**42**](#page42_7b4)

**Example 7(b)-5: Cyclic peptide containing a branched amino acid sequence………………………………………………**[45](#page45_7b5)

***Cross-referenced examples***

Example 3(a)-1: D-amino acids [18](#page18_3a1)

Example 3(c)-1: Enumeration of amino acids by chemical structure [19](#page19_3c1)

Example 3(c)-2: Shorthand formula for an amino acid sequence [20](#page20_3c2)

Example 3(k)-5: Ambiguity abbreviation “Xaa” used in a nonconventional manner [29](#page29_3k5)

Example 27-1: Shorthand formula for a nucleotide sequence [51](#page51_27_1)

Example 27-3: Shorthand formula - four or more specifically defined amino acids [53](#page53_27_3)

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid [56](#page56_29_1)

Example 29-2: Use of the corresponding unmodified amino acid**……….……………………………………….**[57](#page57_29_2)

Example 30-1: Feature key “CARBODHYD” [58](#page58_30_1)

Example 36-1: Sequence with a region of a known number of “X” residues represented

as a single sequence [59](#page59_36_1)

Example 37-1: Sequence with regions of an unknown number of “X” residues must not be represented

as a single sequence [62](#page62_37_1)

Example 37-2: Sequence with regions of an unknown number of “X” residues must not be represented

as a single sequence [63](#page63_37_2)

Example 89-1: Encoding nucleotide sequence and encoded amino acid sequence [65](#page65_89_1)

Example 93-2: Representation of enumerated variants [71](#page71_93_2)

Example 93-3: Representation of a consensus sequence [72](#page72_93_3)

Example 94-1: Representation of single sequence with enumerated alternative amino acids [73](#page73_94_1)

Example 95(a)-1: Representation of a variant sequence by annotation of the primary sequence [74](#page74_95a1)

**Paragraph 8 – Threshold for inclusion of sequences**

***Cross-referenced examples***

Example 3(k)-1: Nucleotide ambiguity symbols [25](#page25_3k1)

Example 3(k)-2: Ambiguity symbol “n” used in both a conventional and nonconventional manner [26](#page26_3k2)

Example 7(a)-1: Branched nucleotide sequence [30](#page30_7a1)

Example 7(a)-6: Nonconventional nucleotide symbols [36](#page36_7a6)

Example 7(b)-1: Four or more specifically defined amino acids [37](#page37_7b1)

Example 7(b)-2: Branched amino acid sequence [38](#page38_7b2)

Example 7(b)-4: Cyclic peptide containing a branched amino acid sequence [42](#page42_7b4)

Example 14-1: The symbol “t” represents uracil in RNA [49](#page49_14_1)

Example 37-1: Sequence with regions of an unknown number of “X” residues must not be represented

as a single sequence [62](#page62_37_1)

Example 37-2: Sequence with regions of an unknown number of “X” residues must not be represented

as a single sequence [63](#page63_37_2)

Example 94-1: Representation of single sequence with enumerated alternative amino acids [73](#page73_94_1)

**Paragraph 11 – Representation of a nucleotide sequence**

***Cross-referenced examples***

Example 3(g)-4: Nucleic Acid Analogues [24](#page24_3g4)

Example 7(a)-1: Branched nucleotide sequence [30](#page30_7a1)

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

**Example 11(a)-1: Double-stranded nucleotide sequence – same lengths** [**46**](#page46_11a1)

**Paragraph 11(b) – Double-stranded nucleotide sequence – not fully complementary**

**Example 11(b)-1: Double-stranded nucleotide sequence – different lengths** [**47**](#page47_11b1)

**Example 11(b)-2: Double-stranded nucleotide sequence – no base-pairing segment** [**48**](#page48_11b2)

**Paragraph 12 – Circular nucleotide sequence**

**Example 12-1: Circular nucleotide sequence <>**

***Cross-referenced examples***

Example 7(a)-1: Branched nucleotide sequence [30](#page30_7a1" \o "page30)

**Paragraph 13 – Representation of nucleotides**

***Cross-referenced examples***

Example 3(k)-2: Ambiguity symbol “n” used in both a conventional and nonconventional manner [26](#page26_3k2)

Example 7(a)-1: Branched nucleotide sequence [30](#page30_7a1)

Example 14-1: The symbol “t” represents uracil in RNA [49](#page49_14_1)

Example 93-1: Representation of enumerated variants [70](#page70_93_1)

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

**Example 14-1: The symbol “t” represents uracil in RNA** [**49**](#page49_14_1)

***Cross-referenced examples***

Example 55-1: Combined DNA/RNA Molecule [64](#page64_55_1)

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

***Cross-referenced examples***

Example 3(g)-1: Nucleotide sequence interrupted by a C3 spacer [21](#page21_3g1)

Example 3(g)-2: Nucleotide sequence with residue alternatives, including a C3 spacer [22](#page22_3g2)

Example 3(k)-4: Ambiguity symbols other than “n” are “specifically defined” [28](#page28_3k4)

Example 95(b)-1: Representation of individual variant sequences with multiple interdependent variations [75](#page75_95b1)

**Paragraph 16 – Representation of a modified nucleotide**

***Cross-referenced examples***

Example 3(g)-1: Nucleotide sequence interrupted by a C3 spacer [21](#page21_3g1)

Example 3(g)-4: Nucleic Acid Analogues [24](#page24_3g4)

**Paragraph 17 – Annotation of a modified nucleotide**

***Cross-referenced examples***

Example 3(g)-1: Nucleotide sequence interrupted by a C3 spacer [21](#page21_3g1)

Example 3(g)-3: Abasic site ...[23](#page23_3g3)

Example 7(a)-1: Branched nucleotide sequence [30](#page30_7a1)

Example 7(a)-2: Linear nucleotide sequence having a secondary structure……………………………………[32](#page32_7a2)

Example 7(a)-6: Nonconventional nucleotide symbols [36](#page36_7a6)

**Paragraph 18 – Annotation of regions of consecutive modified nucleotides**

***Cross-referenced examples***

Example 3(g)-4: Nucleic Acid Analogues [24](#page24_3g4)

Example 11(b)-1: Double-stranded nucleotide sequence – different lengths [47](#page47_11b1)

**Paragraph 19 – Annotation of uracil in DNA or thymine in RNA**

**Cross-referenced examples**

Example 14-1: The symbol “t” represents uracil in RNA [49](#page49_14_1)

**Paragraph 25 – Amino acid sequence residue position number 1**

***Cross-referenced examples***

Example 3(a)-1: D-amino acids [18](#page18_3a1)

Example 7(b)-4: Cyclic peptide containing a branched amino acid sequence [42](#page42_7b4)

Example 7(b)-5: Cyclic peptide containing a branched amino acid sequence [45](#page45_7b5)

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid [56](#page56_29_1)

**Paragraph 26 – Representation of amino acids**

***Cross-referenced examples***

Example 7(b)-2: Branched amino acid sequence [38](#page38_7b2)

Example 7(b)-4: Cyclic peptide containing a branched amino acid sequence [42](#page42_7b4)

Example 7(b)-5: Cyclic peptide containing a branched amino acid sequence ………………………………….[45](#page45_7b5)

Example 36-1: Sequence with a region of a known number of “X” residues represented

as a single sequence [59](#page59_36_1)

Example 89-1: Encoding nucleotide sequence and encoded amino acid sequence [65](#page65_89_1)

Example 92-1: Amino acid sequence encoded by a coding sequence with introns [68](#page68_92_1)

Example 93-2: Representation of enumerated variants [71](#page71_93_2)

Example 93-3: Representation of a consensus sequence [72](#page72_93_3)

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

**Example 27-1: Shorthand formula for an amino acid sequence** [**51**](#page51_27_1)

**Example 27-2: Shorthand formula - less than four specifically defined amino acids** [**52**](#page52_27_2)

**Example 27-3: Shorthand formula - four or more specifically defined amino acids** [**53**](#page53_27_3)

***Cross-referenced examples***

Example 3(c)-2: Shorthand formula for an amino acid sequence……………………………………… [20](#page20_3c2)

Example 7(b)-1: Four or more specifically defined amino acids [37](#page37_7b1)

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid [56](#page56_29_1)

Example 36-1: Sequence with a region of a known number of “X” residues represented

as a single sequence [59](#page59_36_1)

Example 36-2: Sequence with multiple regions of a known number or range of “X” residues

represented as a single sequence [60](#page60_36_2)

Example 36-3: Sequence with multiple regions of a known number or range of “X” residues

represented as a single sequence [61](#page61_36_3)

Example 37-2: Sequence with regions of an unknown number of “X” residues must not be represented as a single sequence………………………………………………………………. [63](#page63_37_2)

Example 93-3: Representation of a consensus sequence [72](#page72_93_3)

Example 94-1: Representation of single sequence with enumerated alternative amino acids [73](#page73_94_1)

Example 95(a)-1: Representation of a variant sequence by annotation of the primary sequence [74](#page74_95a1)

**Paragraph 28 – Amino acid sequences separated by internal terminator symbols**

**Example 28-1: Encoding nucleotide sequence and encoded amino acid sequence** [**54**](#page54_28_1)

***Cross-referenced examples***

Example 89-1: Encoding nucleotide sequence and encoded amino acid sequence [65](#page65_89_1)

Example 92-1: Amino acid sequence encoded by a coding sequence with introns [68](#page68_92_1)

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

**Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid** [**56**](#page56_29_1)

**Example 29-2: Use of the corresponding unmodified amino acid…………………………………………………………….**[**57**](#page57_29_2)

***Cross-referenced examples***

Example 3(a)-1: D-amino acids [18](#page18_3a1)

Example 7(b)-2: Branched amino acid sequence [38](#page38_7b2)

Example 7(b)-3: Branched amino acid sequence………………………………………………………………… [41](#page41_7b3)

Example 7(b)-4: Cyclic peptide containing a branched amino acid sequence………………………………… [42](#page42_7b4)

Example 30-1: Feature key “CARBODHYD” [58](#page58_30_1)

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

**Example 30-1: Feature key “CARBODHYD”** [**58**](#page58_30_1)

***Cross-referenced examples***

Example 3(a)-1: D-amino acids [18](#page18_3a1)

Example 3(c)-1: Enumeration of amino acids by chemical structure [19](#page19_3c1)

Example 7(b)-2: Branched amino acid sequence [38](#page38_7b2)

Example 7(b)-3: Branched amino acid sequence [41](#page41_7b3)

Example 7(b)-4: Cyclic peptide containing a branched amino acid sequence………………………………… [42](#page42_7b4)

Example 7(b)-5: Cyclic peptide containing a branched amino acid sequence………………………………… [45](#page45_7b5)

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid…………………………………. [56](#page56_29_1)

Example 29-2: Use of the corresponding unmodified amino acid…………………………………..……………[57](#page57_29_2)

**Example 30-2: Post-translationally modified amino acids <>**

***Cross-referenced examples***

Example 3(a)-1: D-amino acids [18](#page18_3a1" \o "page18)

Example 7(b)-2: Branched amino acid sequence [38](#page38_7b2" \o "page38)

Example 7(b)-3: Branched amino acid sequence [41](#page41_7b3" \o "page41)

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

***Cross-referenced examples***

Example 3(a)-1: D-amino acids [18](#page18_3a1)

Example 3(c)-1: Enumeration of amino acids by chemical structure [19](#page19_3c1)

Example 7(b)-2: Branched amino acid sequence [38](#page38_7b2)

Example 7(b)-3: Branched amino acid sequence [41](#page41_7b3)

Example 7(b)-4: Cyclic peptide containing a branched amino acid sequence………………………………… [42](#page42_7b4)

Example 7(b)-5: Cyclic peptide containing a branched amino acid sequence………………………………… [45](#page45_7b5)

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

***Cross-referenced examples***

Example 3(c)-1: Enumeration of amino acids by chemical structure [19](#page19_3c1)

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

***Cross-referenced examples***

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid [56](#page56_29_1)

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

**Example 36-1: Sequence with a region of a known number of “X” residues represented as a single sequence** [**59**](#page59_36_1)

**Example 36-2: Sequence with multiple regions of a known number or range of “X” residues**

**represented as a single sequence** [**60**](#page60_36_2)

**Example 36-3: Sequence with multiple regions of a known number or range of “X” residues**

**represented as a single sequence** [**61**](#page61_36_3)

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

**Example 37-1: Sequence with regions of an unknown number of “X” residues must not be represented**

**as a single sequence** [**62**](#page62_37_1)

**Example 37-2: Sequence with regions of an unknown number of “X” residues must not be represented**

**as a single sequence** [**63**](#page63_37_2)

**Paragraph 41 – Reserved characters**

***Cross-referenced examples***

Example 89-2: Feature location extends beyond the disclosed sequence [66](#page66_89_2" \o "page66)

**Paragraph 54 – The element INSDSeq\_moltype**

***Cross-referenced examples***

Example 14-1: The symbol “t” represents uracil in RNA [49](#page49_14_1)

**Paragraph 55 – A nucleotide sequence that contains both DNA and RNA segments**

**Example 55-1: Combined DNA/RNA Molecule……………………………………………………………………………………**[64](#page64_55_1)

**Paragraph 56 – Example illustrating a nucleotide sequence that contains both DNA and RNA segments**

***Cross-referenced examples***

Example 55-1: Combined DNA/RNA Molecule [64](#page64_55_1)

**Paragraph 57 – The element INSDSeq\_sequence**

***Cross-referenced examples***

Example 28-1: Encoding nucleotide sequence and encoded amino acid sequence [54](#page54_28_1)

Example 92-1: Amino acid sequence encoded by a coding sequence with introns [68](#page68_92_1)

**Paragraph 65 – Location descriptor**

***Cross-referenced examples***

Example 3(g)-4: Nucleic Acid Analogues [24](#page24_3g4)

Example 89-2: Feature location extends beyond the disclosed sequence [66](#page66_89_2)

**Paragraph 66 – Location descriptor syntax**

***Cross-referenced examples***

Example 3(g)-4: Nucleic Acid Analogues [24](#page24_3g4)

Example 7(b)-4: Cyclic peptide containing a branched amino acid sequence [42](#page42_7b4)

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid [56](#page56_29_1)

Example 30-1: Feature key “CARBODHYD” [58](#page58_30_1)

Example 89-2: Feature location extends beyond the disclosed sequence [66](#page66_89_2)

**Paragraph 67 – Location operator**

***Cross-referenced examples***

Example 92-1: Amino acid sequence encoded by a coding sequence with introns [68](#page68_92_1)

**Paragraph 70 – Feature locations**

***Cross-referenced examples***

Example 7(b)-4: Cyclic peptide containing a branched amino acid sequence [42](#page42_7b4)

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid [56](#page56_29_1)

Example 30-1: Feature key “CARBODHYD” [58](#page58_30_1)

Example 89-2: Feature location extends beyond the disclosed sequence [66](#page66_89_2)

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

***Cross-referenced examples***

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid [56](#page56_29_1)

Example 89-2: Feature location extends beyond the disclosed sequence [66](#page66_89_2)

**Paragraph 83 – Example illustrating a nucleotide sequence that is not naturally occurring**

***Cross-referenced examples***

Example 55-1: Combined DNA/RNA Molecule [64](#page64_55_1)

**Paragraph 89 – “CDS” Feature key**

**Example 89-1: Encoding nucleotide sequence and encoded amino acid sequence** [**65**](#page65_89_1)

**Example 89-2: Feature location extends beyond the disclosed sequence** [**66**](#page66_89_2)

***Cross-referenced examples***

Example 92-1: Amino acid sequence encoded by a coding sequence with introns [68](#page68_92_1)

**Paragraph 90 – The qualifiers “transl\_table” and “translation”**

***Cross-referenced examples***

Example 28-1: Encoding nucleotide sequence and encoded amino acid sequence [54](#page54_28_1)

Example 89-1: Encoding nucleotide sequence and encoded amino acid sequence [65](#page65_89_1)

Example 92-1: Amino acid sequence encoded by a coding sequence with introns [68](#page68_92_1)

**Paragraph 92 – Amino acid sequence encoded by a coding sequence**

**Example 92-1: Amino acid sequence encoded by a coding sequence with introns** [**68**](#page68_92_1)

***Cross-referenced examples***

Example 28-1: Encoding nucleotide sequence and encoded amino acid sequence [54](#page54_28_1)

Example 89-1: Encoding nucleotide sequence and encoded amino acid sequence [65](#page65_89_1)

Example 89-2: Feature location extends beyond the disclosed sequence [66](#page66_89_2)

**Paragraph 93 – Primary sequence and a variant, each enumerated by its residue**

**Example 93-1: Representation of enumerated variants** [**70**](#page70_93_1)

**Example 93-2: Representation of enumerated variants** [**71**](#page71_93_2)

**Example 93-3: Representation of a consensus sequence** [**72**](#page72_93_3)

**Paragraph 94 – Variant sequence disclosed as a single sequence with enumerated alternative residues**

**Example 94-1: Representation of single sequence with enumerated alternative amino acids** [**73**](#page73_94_1)

**Example 94-2: Representation of single sequence with enumerated alternative amino acids that may be modified amino acids <>**

***Cross referenced examples***

Example 3(a)-1: D-amino acids 18

Example 7(b)-1: four or more specifically defined amino acids 37

Example 27-1: Shorthand formula for an amino acid 51

Example 27-3: Shorthand formula - four or more specifically defined amino acids 53

Example 29-1: Most restrictive ambiguity symbol for an "other" amino acid 56**Paragraph 95(a) – A variant sequence disclosed only by reference to a primary sequence with multiple independent variations**

**Example 95(a)-1: Representation of a variant sequence by annotation of the primary sequence** [**74**](#page74_95a1)

**Paragraph 95(b) – A variant sequence disclosed only by reference to a primary sequence with multiple interdependent variations**

**Example 95(b)-1: Representation of individual variant sequences with multiple interdependent variations** [**75**](#page75_95b1)

**Paragraph 96 – Feature keys and qualifiers for a variant sequence**

***Cross-referenced examples***

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid [56](#page56_29_1)

**Paragraph 97 – Annotation of a variant sequence**

***Cross-referenced examples***

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid [56](#page56_29_1)

Example 93-3: Representation of a consensus sequence [72](#page72_93_3)

Example 94-1: Representation of single sequence with enumerated alternative amino acids [73](#page73_94_1)

## EXAMPLES

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

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

A patent application describes the following sequence:

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

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

**YES**

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

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

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

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

Paragraph 25 indicates that when amino acid sequences are circular in configuration and the ring consists solely of amino acid residues linked by peptide bonds, applicant must choose the amino acid in residue position number 1. Accordingly, the sequence may be represented as:

AEKXGMX (SEQ ID NO: 1)

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

**Relevant ST.26 paragraphs: 3(a)**, 7(b), 25, 26, 29, 30, and 31

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

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



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

**YES**

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

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

The sequence may be represented as:

VAFXGK (SEQ ID NO: 2)

wherein “X” represents an “other” modified amino acid:, which requires a feature key “SITE” together with the qualifier “note”. The qualifier “note” provides the complete, unabbreviated name of the modified tryptophan in position 4 of the enumerated peptide, e.g., “6-amino-7-(1H-indol-3-yl)-5-oxoheptanoic acid”. The methylation of the C terminus changes the chemical structure of the terminal lysine since the -OH on the terminal end is replaced by -CH3. Due to this structural change, the lysine within the sequence is considered a “modified amnio acid.” Accordingly, a feature key “SITE” and qualifier “note” are required to indicate the methylation of the C-terminus. Valine, however, is not considered a “modified amino acid” since the addition of the acetyl group to the valine involves a conventional peptide linkage. The acetylation does not alter the structure of the valine. Accordingly, an additional feature key “SITE” and qualifier “note” should be included to indicate the acetylation of the N terminus.

Alternatively, the sequence may be represented as:

VAFW (SEQ ID NO: 3)

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

**Relevant ST.26 paragraph(s): 3(c)**, 7(b), 29, 30, and 31

**Example 3(c)-2: Shorthand formula for an amino acid sequence**

(G4z)n

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

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

**YES**

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

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

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

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

GGGGX (SEQ ID NO: 4)

and should be annotated with the feature key REGION, feature location “&gt;5” (corresponds to >5), with a note qualifier with the value “The entire sequence of amino acids 1-5 can be repeated one or more times.”

According to paragraph 27, “” 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. Since in this example “X” represents “any amino acid”, it must be annotated with the feature key VARIANT and a note qualifier with the value, “X can be any amino acid”.

Where practicable, each “X” should be annotated individually. However, a region of contiguous “X” residues, or a multitude of “X” residues dispersed throughout the sequence, may be jointly described with the feature key VARIANT using the syntax “x..y” as the location descriptor, where x and y are the positions of the first and last “X” residues, and a note qualifier with the value, “X can be any amino acid”.

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

**Relevant ST.26 paragraph(s): 3(c)**, 7(b) and 27.

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

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

A patent application describes the following sequence:

atgcatgcatgcncggcatgcatgc

where n = a C3 spacer with the following structure:



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

**YES**

The enumerated sequence contains two segments of specifically defined nucleotides separated by a C3 spacer.

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

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

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

atgcatgcatgc (SEQ ID NO: 5)

cggcatgcatgc (SEQ ID NO: 6)

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

**Relevant ST.26 paragraphs: 3(g)**, 7(a), and 15

**Example 3(g)-2: Nucleotide sequence with residue alternatives, including a C3 spacer**

A patent application describes the following sequence:

atgcatgcatgcncggcatgcatgc

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



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

**YES**

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

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

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

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

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

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

atgcatgcatgcvcggcatgcatgc (SEQ ID NO: 7)

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

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

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

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

**Relevant ST.26 paragraphs: 3(g)**, 7(a), and 15

**Example 3(g)-3: Abasic site**

A patent application describes the following sequence:

gagcattgac-AP-taaggct

Wherein AP is an abasic site

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

**YES**

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

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

The sequence must be included in a sequence listing as:

gagcattgacntaaggct (SEQ ID NO: 10)

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

**Relevant ST.26 paragraphs: 3(g)**, 7(a), and 17

**Example 3(g)-4: Nucleic Acid Analogues**

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

PO4-tagttcattgactaaggctccccattgact-OH

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

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

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

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

GNA sequences do not have a 5’-end and a 3’-end, but rather, a 3’-end and a 2’-end. The 3’-end, which is routinely depicted as having a terminal phosphate group, corresponds to the 5’-end of DNA or RNA. (Note that other nucleic acid analogues may correspond differently to the 5’-end and 3’-end of DNA and RNA.) According to paragraph 11, it must be included in a sequence listing “in the direction from left to right that mimics the 5’-end to 3’-end direction.” Therefore, it must be included in a sequence listing as:

tagttcattgactaaggctccccattgact (SEQ ID NO: 11)

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

**Relevant ST.26 paragraphs:** 3(d), **3(g)**, 7(a), 11, 16, 18, 65, and 66

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

**Example 3(k)-1: Nucleotide ambiguity symbols**

5’ NNG KNG KNG K 3’

N and K are IUPAC-IUB ambiguity codes

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

**NO**

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

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

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

**NO**

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

**Relevant ST.26 paragraphs: 3(k)**, 7(a), 8, and 13

**Example 3(k)-2: Ambiguity symbol “n” used in both a conventional and nonconventional manner**

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

(i) in one embodiment, N is any nucleotide;

(ii) in one embodiment, N is optional but is preferably G;

(iii) in one embodiment, N is K;

(iv) in one embodiment, N is C.

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

**NO**

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

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

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

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

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

**NO**

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

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

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

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

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

aatgccggag (SEQ ID NO: 12)

aatgccggak (SEQ ID NO: 13)

aatgccggac (SEQ ID NO: 14)

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

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

**Relevant ST.26 paragraphs: 3(k)**, 7(a), 8, and 13

**Example 3(k)-3: Ambiguity symbol “n” used in a nonconventional manner**

An application discloses the sequence: 5’-aatgttggan-3’

Wherein n is c

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

**YES**

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

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

5’-aatgttggac-3’

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

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

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

**Relevant ST.26 paragraphs: 3(k)** and 7(a)

**Example 3(k)-4: Ambiguity symbols other than “n” are “specifically defined”**

A patent application describes the following sequence:

5’ NNG KNG KNG KAG VCR 3’

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

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

**YES**

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

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

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

The sequence must be included in a sequence listing as:

nngkngkngkagvcr (SEQ ID NO: 16)

**Relevant ST.26 paragraphs: 3(k)**, 7(a) and 15

**Example 3(k)-5: Ambiguity abbreviation “Xaa” used in a nonconventional manner**

A patent application describes the following sequence:

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

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

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

**YES**

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

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

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

XYEKGJL (SEQ ID NO: 17)

According to paragraph 27, “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. Since “X” at position 1 of SEQ ID NO: 17 represents “any amino acid”, it must be annotated with the feature key VARIANT and a note qualifier with the value, “X can be any amino acid”.

Where practicable, each “X” should be annotated individually. However, a region of contiguous “X” residues, or a multitude of “X” residues dispersed throughout the sequence, may be jointly described with the feature key VARIANT using the syntax “x..y” as the location descriptor, where x and y are the positions of the first and last “X” residues, and a note qualifier with the value, “X can be any amino acid”.

**Relevant ST.26 paragraphs: 3(k)**, 7(b), 26, and 27

### *Paragraph 7(a) – Nucleotide sequences required in a sequence listing*

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

The description discloses the following branched nucleotide sequence:



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

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

SP(O-)(=O)CACACAAAAAAAAAAAAAAAAAAAAAAAAA 3’ is segments B, C, and D; and

SP(O-)(=O)CACATAGGCATCTCCTAGTGCAGGAAGA 3’ is segment E.

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

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

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

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

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

The four vertical segments B-E each contain more than ten specifically defined nucleotides, wherein adjacent nucleotides are joined 3’ to 5’, and therefore each is required to be included in a sequence listing.

In horizontal segment A, the linear regions of the nucleotide sequence are linked by the non-nucleotide moiety “pnp” and each of these linked linear regions contains fewer than ten specifically defined nucleotides. Therefore, since no region of segment A contains ten or more specifically defined nucleotides wherein adjacent nucleotides are joined 3’ to 5’, they are not required by ST.26 paragraph 7(a) to be included in a sequence listing.

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

**NO**

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

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

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

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

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

cacacaaaaaaaaaaaaaaaaaaaaaaaaa (SEQ ID NO: 18)

The first “c” in the sequence should be further described using the feature key “misc\_feature” and the qualifier “note” with the value e.g., “This sequence is one of four branches of a branched polynucleotide”.

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

cacataggcatctcctagtgcaggaaga (SEQ ID NO: 19)

The first “c” in the sequence should be further described using the feature key “misc\_feature” and the qualifier “note” with the value e.g., “This sequence is one of four branches of a branched polynucleotide.”

**Relevant ST.26 paragraph(s): 7(a)**, 8, 11, 13, and 17

**Example 7(a)-2: Linear nucleotide sequence having a secondary structure**

A patent application describes the following sequence**:**



Wherein Ψ is pseudouridine.

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

**YES**

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

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

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

The symbol “u” must not be used to represent uracil in an RNA molecule in the sequence listing. According to paragraph 14, the symbol “t” will be construed as uracil in RNA. The sequence must be included as:

gcggatttagctcagctgggagagcgccagactgaatanctggagtcctgtgtncgatccacagaattcgcacca (SEQ ID NO: 20)

The value of the mandatory “mol\_type” qualifier of the mandatory “source” feature key is “tRNA”. Additional information may be provided with feature key “tRNA” and any appropriate qualifier(s).

The “n” residues must be further described in a feature table using the feature key “modified\_base” and the mandatory qualifier “mod\_base” with the abbreviation “p” for pseudouridine as the qualifier value (see Annex 1, Table 2).

**Relevant ST.26 paragraph(s): 7(a)**, 11, 13, 14, 17, 62, 84 and Annex I, sections 2 and 5, feature key 5.43

**Example 7(a)-3: Nucleotide ambiguity symbols used in a nonconventional manner**

A patent application describes the following sequence:

5’ GATC-MDR-MDR-MDR-MDR-GTAC 3’

The explanation of the sequence in the disclosure further indicates: “A “DR Element” consists of the sequence 5’ ATCAGCCAT 3’. A mutant DR Element, or MDR, is a DR element wherein the middle 5 nucleotides, CAGCC, are mutated to TTTTT.”

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

**YES**

The enumerated sequence uses the symbol “MDR”. Where it is unclear if a symbol used in a sequence is intended to be a conventional symbol, i.e., a symbol set forth in Annex 1, Section 3, Table 3, or a nonconventional symbol, the explanation of the sequence in the disclosure must be consulted to make a determination (see Introduction to this document). According to Table 3, “MDR” could be interpreted as three conventional symbols (m = a or c, d = a or g or t/u, r = g or a) or as an abbreviation that is short-hand notation for some other structure.

Consultation of the disclosure indicates that an MDR element is equivalent to 5’ ATTTTTTAT 3’. The letters “MDR” are considered conventional symbols used in a nonconventional manner; therefore, the sequence must be interpreted as though it were disclosed using the equivalent conventional symbols. Accordingly, the enumerated sequence that is considered for inclusion in a sequence listing is:

5’ GATC ATTTTTTAT ATTTTTTAT ATTTTTTAT ATTTTTTAT GTAC 3’

The enumerated sequence has 44 specifically defined nucleotides and is required by ST.26 paragraph 7(a) to be included in a sequence listing.

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

The sequence must be included in a sequence listing as:

gatcattttttatattttttatattttttatattttttatgtac (SEQ ID NO: 21)

**Relevant ST.26 paragraphs: 7(a)** and 13

**Example 7(a)-4: Nucleotide ambiguity symbols used in a nonconventional manner**

A patent application describes the following sequence:

5’ ATTC-N-N-N-N-GTAC 3’

The explanation of the sequence in the disclosure further indicates that “N” consists of the sequence 5’ ATACGCACT 3’.

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

**YES**

The enumerated sequence uses the symbol “N”. The explanation of the sequence in the disclosure must be consulted to determine if the “N” is used in a conventional or nonconventional manner (see Introduction to this document).

Consultation of the disclosure indicates that “N” is equivalent to 5’ ATACGCACT 3’. Thus, the “N” is a conventional symbol used in a nonconventional manner. Accordingly, the sequence must be interpreted as though it were disclosed using the equivalent conventional symbols:

5’ ATTC-ATACGCACT-ATACGCACT-ATACGCACT-ATACGCACT-GTAC 3’

The enumerated sequence has 44 specifically defined nucleotides and is required by ST.26 paragraph 7(a) to be included in a sequence listing.

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

The sequence must be included in a sequence listing as:

attcatacgcactatacgcactatacgcactatacgcactgtac (SEQ ID NO: 22)

**Relevant ST.26 paragraphs: 7(a)** and 13

**Example 7(a)-5: Nonconventional nucleotide symbols**

A patent application describes the following sequence:

5’ GATC-β-β-β-β-GTAC 3’

The explanation of the sequence in the disclosure further indicates that “β” consists of the sequence 5’ ATACGCACT 3’.

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

**YES**

The enumerated sequence uses the nonconventional symbol “β”. The explanation of the sequence in the disclosure must be consulted to determine the meaning of “β” (see Introduction to this document).

Consultation of the disclosure indicates that “β” is equivalent to 5’ ATACGCACT 3’. Thus, the “β” is a nonconventional symbol used to represent a sequence of nine specifically defined, conventional symbols. Accordingly, the sequence must be interpreted as though it were disclosed using the equivalent conventional symbols:

5’ GATC-ATACGCACT-ATACGCACT-ATACGCACT-ATACGCACT-GTAC 3’

The enumerated sequence has 44 specifically defined nucleotides and is required by ST.26 paragraph 7(a) to be included in a sequence listing.

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

The sequence must be included in a sequence listing as:

gatcatacgcactatacgcactatacgcactatacgcactgtac (SEQ ID NO: 23)

**Relevant ST.26 paragraphs: 7(a)** and 13

**Example 7(a)-6: Nonconventional nucleotide symbols**

A patent application describes the following sequence:

5’ GATC-β-β-β-β-GTAC 3’

The explanation of the sequence in the disclosure further indicates that “β” is equal to adenine, inosine, or pseudouridine.

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

**NO**

The enumerated sequence uses the nonconventional symbol “β”. The explanation of the sequence in the disclosure must be consulted to determine the meaning of “β” (see Introduction to this document).

Consultation of the disclosure indicates that “β” is equivalent to adenine, inosine, or pseudouridine. The only conventional symbol that can be used to represent “adenine, inosine, or pseudouridine” is “n”; therefore, the “β” is a nonconventional symbol used to represent the conventional symbol “n”. Accordingly, the sequence must be interpreted to have four “n” symbols (shown as “N” below) in place of the four “β” symbols:

5’ GATC-N-N-N-N-GTAC 3’

The enumerated sequence has only eight specifically defined nucleotides and is not required by ST.26 paragraph 7(a) to be included in a sequence listing.

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

**NO**

The enumerated sequence, 5’ GATC-N-N-N-N-GTAC 3’ must not be included in a sequence listing.

However, a disclosed alternative sequence may be included in a sequence listing if at least 2 of the “n” symbols are replaced by adenine, resulting in a sequence with at least 10 or more specifically defined nucleotides.

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

One possible permitted representation is:

gatcaaaagtac (SEQ ID NO: 24)

In the above example, the four adenine nucleotides that replace the β symbols should be annotated to note that these positions could be substituted with inosine or pseudouridine.

The feature key “misc\_difference” should be used with a feature location 5-8 and a qualifier “note” with the value, e.g., “A nucleotide in any of positions 5-8 may be replaced with inosine or pseudouridine”. Since these alternatives are modified nucleotides, then the feature key “modified\_base” together with the qualifier “mod\_base” would be required. The value for the “mod\_base” qualifier can be “OTHER” with a “note” qualifier and the value of “i or p”.

Other permutations are possible.

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

**Relevant ST.26 paragraphs: 7(a)**, 8, 13, and 17

### *Paragraph 7(b) – Amino Acid sequences required in a sequence listing*

**Example 7(b)-1: Four or more specifically defined amino acids**

XXXXXXXXDXXXXXXXXXXFXXXXXXXXXXXXXXXXXXXXXXXXXXXXAXXXXXXXXXXXXXXXXXXXGXXXXX

Where X = any amino acid

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

**YES**

The enumerated peptide contains four specifically defined amino acids. The symbol “X” is used conventionally to represent the remaining amino acids as any amino acid (see Introduction to this document).

Because there are four specifically defined amino acids, i.e., Asp, Phe, Ala and Gly, ST.26 paragraph 7(b) requires that the sequence be included in a sequence listing.

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

The sequence must be represented as:

XXXXXXXXDXXXXXXXXXXFXXXXXXXXXXXXXXXXXXXXXXXXXXXXAXXXXXXXXXXXXXXXXXXXGXXXXX (SEQ ID NO: 25)

According to paragraph 27, “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. Since “X” in SEQ ID NO: 25 represents “any amino acid”, it must be annotated with the feature key VARIANT and a note qualifier with the value, “X can be any amino acid”.

Where practicable, each “X” should be annotated individually. However, a region of contiguous “X” residues, or a multitude of “X” residues dispersed throughout the sequence, may be jointly described with the feature key VARIANT using the syntax “x..y” as the location descriptor, where x and y are the positions of the first and last “X” residues, and a note qualifier with the value, “X can be any amino acid”.

**Relevant ST.26 paragraph(s): 7(b)**, 8 and 27

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

The application describes a branched sequence where the Lysine residues are used as a scaffolding core to form eight branches to which multiple linear peptide chains are attached. Lysine is a dibasic amino acid, providing it with two sites for peptide-bonding. The peptide is illustrated as follows:



In the above branched peptide, the bonds between lysine and another amino acid depicted by represent an amide linkage between the terminal amine of the lysine and the carboxyl end of the bonded amino acid. The bonds depicted by represent an amide linkage between the side chain amine of the lysine and the carboxyl end the bonded amino acid.

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

**YES**

The example discloses a branched sequence where the lysine residues are used as a scaffolding. Paragraph 7(b) requires that the unbranched or linear region of the sequence, containing four or more specifically defined amino acids, be included in a sequence listing. In the above example, the linear regions of the branched peptide that have four or more specifically defined amino acids are encircled:



ST.26 paragraph 7(b) requires inclusion of peptides 1-6 above in a sequence listing.

Peptides which are not required to be included in the sequence listing are:

YFA

LLK

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

**NO**

According to paragraph 8, a sequence listing must not include any sequences having fewer than four specifically defined amino acids.

The peptides YFA and LLK each contain only three specifically defined amino acids and therefore, they must not be included in a sequence listing as separate sequences with their own sequence identification numbers.

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

Peptides 1-6 must be represented with separate sequence identifiers:

RISL (SEQ ID NO: 26)

LLKK (SEQ ID NO: 27)

IPACTA (SEQ ID NO: 28)

FRAGGK (SEQ ID NO: 29)

HQYFA (SEQ ID NO: 30)

ATFGKKKA (SEQ ID NO: 31)

The branched structure may be annotated using the feature key “SITE” and the mandatory qualifier “note” with the value e.g., “This sequence is one part of a branched amino acid sequence”. According to ST.26 paragraph 30, SEQ ID Nos 27, 29, and 31, must include an annotation for each lysine to indicate that it is a modified amino acid, using the feature key “SITE” together with the qualifier “note” describing that the side chain of the lysine is linked via an amide linkage to another sequence. Each of the SEQ ID Nos 26, 28, and 30 should include an annotation to indicate that the C-terminal amino acid is linked to another sequence, using the feature key “SITE” together with the qualifier “note”.

**Relevant ST.26 paragraph(s): 7(b)**, 8, 26, 29, 30, and 31

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

Peptide of the following sequence:



The linkage between the terminal Glycine residue in the lower sequence and the Lysine in the upper sequence is through an amide bond between the carboxy terminus of the Glycine and the amino terminal side chain of the Lysine.

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

**YES**

The unbranched or linear region of a sequence, containing four or more specifically defined amino acids, must be included in a sequence listing. In the above example, the linear regions of the branched peptide that have more than four amino acids are:



ST.26 paragraph 7(b) requires inclusion of sequences 1 and 2 in a sequence listing.

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

Sequences 1 and 2 must be represented with separate sequence identifiers:

DGSAKKKK (SEQ ID NO: 32)

AASHG (SEQ ID NO: 33)

The sequence DGSAKKKK must include an annotation to indicate that the lysine in position number 5 is a modified amino acid, using the feature key “SITE” together with the qualifier “note” describing that the side chain of the lysine is linked via an amide linkage to another sequence. The sequence AASHG should include an annotation to indicate that the glycine in position number 5 is linked to another sequence using the feature key “SITE” together with the qualifier “note”.

**Relevant ST.26 paragraph(s): 7(b)**, 26, 29, 30, and 31

**Example 7(b)-4: Cyclic peptide containing a branched amino acid sequence**

A patent application discloses the following structure:



The Cysteine and Leucine in the cyclic structure are linked through the side chain of the Cys and carboxy terminus of the Leu.

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

The structure shown is a branched cyclic amino acid sequence which contains the following amino acids:



Since the side chain of the Cys and carboxy terminus of the Leu are involved in the cyclization, the N-terminus of the cyclic peptide is located at Cys-1.

**YES – the cyclic region of the peptide**

ST.26 paragraph 7(b) requires that the linear region of a branched sequence containing four or more specifically defined amino acids, wherein the amino acids form a single peptide backbone, must be included in a sequence listing. In the above example, the cyclic region of the branched peptide has more than four amino acids, and therefore, must be included in a sequence listing.

**NO – the tripeptide branch of the peptide**

The tripeptide branch Ala-Leu-Glu is not required to be in the sequence listing.

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

**NO**

According to paragraph 8, a sequence listing must not include any sequences having fewer than four specifically defined amino acids.

The tripeptide branch contains only three specifically defined amino acids and therefore, it must not be included in a sequence listing as a separate sequence with its own sequence identification number.

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

While this example illustrates a peptide that is circular in configuration, the ring does not consist solely of amino acid residues in peptide linkages, as indicated in paragraph 25. Since the cyclization of the amino acid sequence occurs through the side chain of cysteine (Cys) and the carboxy terminus of the leucine (Leu), the cysteine must be assigned position number 1 within the cyclic region of the peptide. Accordingly, the sequence must be represented as:

CALRDKL (SEQ ID NO: 90)

As indicated in the figure above, the amino acid sequence is cyclized through a thioester conjugation between the cysteine side chain and the carboxy terminus of the leucine. The feature key “SITE” must be used to describe the modified cysteine, which forms the intrachain linkage with leucine. The feature location element is the residue numbers of the cross-linked amino acids in “x..y” format, i.e., “1..7”. The mandatory qualifier “note” should indicate the nature of the linkage, e.g., “cysteine leucine thioester (Cys-Leu)”, to specify that Cys-1 and Leu-7 are linked through a thioester bond. Further, the lysine in position number 6 must be annotated to indicate that it is modified, by using the feature key “SITE” together with the mandatory qualifier “note”, where the qualifier value describes that the lysine side chain links the tripeptide ALE.

**Relevant ST.26 paragraphs: 7(b),** 8, 25, 26, 29, 30, 31, 66(c), and 70

**Example 7(b)-5: Cyclic peptide containing a branched amino acid sequence**

A patent application discloses the following branched cyclic peptide:



The Ser and the Lys are linked through an amide bond between the carboxy terminus of the serine and amine in the side chain of the Lys.

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

**YES**

Paragraph 7(b) requires inclusion of any sequence that contains four or more specifically defined amino acids and which can be represented as a linear region of a branched sequence in a sequence listing. In the above example, the peptide contains a cyclic region wherein the amino acids are joined by peptide bonds, and a branched region which is joined to a side chain of the Lys in the cyclic region. The regions of this branched peptide which can be represented as linear and which contain four or more specifically defined amino acids are:



ST.26 requires inclusion of sequences 1 and 2 of this cyclic branched peptide in a sequence listing, each with their own sequence identification number.

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

Sequence 1 must be represented as:

LRDQS (SEQ. ID. NO: 91)

Sequence 1 may be annotated by using the feature key “SITE” together with the qualifier “note” to describe that the serine in position 5 is linked to another sequence through an amide linkage between Ser and a side chain of a Lys in the other sequence.

Sequence 2 is a cyclic peptide. Paragraph 25 indicates that when an amino acid sequence is circular in configuration and has no amino and carboxy termini, applicant must choose the amino acid residue in position number 1. Accordingly, the sequence may be represented as:

ALFKNG (SEQ. ID. NO: 92)

Alternatively, any other amino acid in the sequence could be designated as residue position number 1. The sequence ALFKNG must be further described using the feature key “SITE” together with the qualifier “note” to describe that the side chain of the Lys in residue position number 4 is linked via an amide linkage to another sequence. This side chain linkage modifies the Lys, and according to ST.26 paragraph 30, a modified amino acid must be further described in the feature table. Moreover, a feature key “REGION” and a qualifier “note” should be provided to indicate that the peptide ALFKNG is circular.

**Relevant ST.26 paragraphs: 7(b)**, 25, 26, 30, and 31

### 

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

**Example 11(a)-1: Double-stranded nucleotide sequence – same lengths**

A patent application describes the following double-stranded DNA sequence:

3’-CCGGTTAACGCTA-5’

5’-GGCCAATTGCGAT-3’

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

**YES**

Each enumerated nucleotide sequence has more than 10 specifically defined nucleotides. At least one strand must be included in the sequence listing, because the two strands of this double-stranded nucleotide sequence are fully complementary to each other.

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

**YES**

While the sequence of only one strand must be included in the sequence listing, the sequences of both strands may be included, each with its own sequence identification number.

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

The double-stranded DNA sequence must be represented either as a single sequence or as two separate sequences. Each sequence included in the sequence listing must be represented in the 5’ to 3’ direction and assigned its own sequence identification number.

atcgcaattggcc (top strand) (SEQ ID NO: 34)

and/or

ggccaattgcgat (bottom strand) (SEQ ID NO: 35)

**Relevant ST.26 paragraphs:** 7(a), **11(a)**, and 13

### 

### *Paragraph 11(b) – Double-stranded nucleotide sequence - not fully complementary*

**Example 11(b)-1: Double-stranded nucleotide sequence – different lengths**

A patent application contains the following drawing and caption:

5’-tagttcattgactaaggctccccattgactaaggcgactagcattgactaaggcaagc-3’

||||||||||||||||

gggtaactgantccgc

The human gene ABC1 promoter region (top strand) bound by a PNA probe (bottom strand), where “n” in the PNA probe is a universal PNA base selected from the group consisting of 5-nitroindole and 3-nitroindole.

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

**YES** – theABC1 promoter region (top strand)

The top strand has more than ten enumerated and “specifically defined” nucleotides and is required to be included in a sequence listing.

**YES** – thePNA probe (bottom strand)

The bottom strand must also be included in the sequence listing, with its own sequence identification number, because the two strands are not fully complementary to each other. The individual residues that comprise a PNA or “peptide nucleic acid” are considered nucleotides according to ST.26 paragraph 3(g). Therefore, the bottom strand has more than 10 enumerated and “specifically defined” nucleotides and is required to be included in a sequence listing.

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

The top strand must be included in a sequence listing as:

tagttcattgactaaggctccccattgactaaggcgactagcattgactaaggcaagc (SEQ ID NO: 36)

The bottom strand is a peptide nucleic acid and therefore does not have a 3’ and 5’ end. According to paragraph 11, it must be included in a sequence listing “in the direction from left to right that mimics the 5’–end to 3’-end direction.” Therefore, it must be included in a sequence listing as:

cgcctnagtcaatggg (SEQ ID NO: 37)

The “organism” qualifier of the feature key “source” must have the value “synthetic construct” and the mandatory qualifier “mol\_type” with the value “other DNA”. The bottom strand must be described in a feature table using the feature key “modified\_base” and the mandatory qualifier “mod\_base” with the abbreviation “OTHER”. A “note” qualifier must be included with the complete unabbreviated name of the modified nucleotides, such as “N-(2-aminoethyl) glycine nucleosides”.

The “n” residue must be further described in a feature table using the feature key “modified\_base” and the mandatory qualifier “mod\_base” with the abbreviation “OTHER”. A “note” qualifier must be included with the complete unabbreviated name of the modified nucleotide: “N-(2-aminoethyl) glycine 5-nitroindole or N-(2-aminoethyl) glycine 3-nitroindole”.

**Relevant ST.26 paragraphs:** 3(g), 7(a), **11(b)**, 17, and 18

**Example 11(b)-2: Double-stranded nucleotide sequence – no base-pairing segment**

A patent application describes the following double-stranded DNA sequence:

3’-CCGGTTAGCTTATACGCTAGGGCTA-5’

||||||| ||||||||||||

5’-GGCCAATATGGCTTGCGATCCCGAT-3’

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

**YES**

Each strand of the enumerated, double-stranded nucleotide sequence has more than 10 specifically defined nucleotides. Both strands must be included in the sequence listing, each with its own sequence identification number, because the two strands are not fully complementary to each other.

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

The sequence of each strand must be represented in the 5’ to 3’ direction and assigned its own sequence identification number:

atcgggatcgcatattcgattggcc (top strand) (SEQ ID NO: 38)

and

ggccaatatggcttgcgatcccgat (bottom strand) (SEQ ID NO: 39)

**Relevant ST.26 paragraphs:** 7(a), **11(b)**, and 13

*Paragraph 12 – Circular nucleotide sequence*

**Example 12-1: Circular nucleotide sequence**

A patent application contains the following figure, disclosing the DNA sequence of plasmid pCIRC1:



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

**YES**

The enumerated nucleotide sequence has more than 10 specifically defined nucleotides. Therefore, the sequence must be included in a sequence listing as required by ST.26 paragraph (7)(a).

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

According to ST.26 paragraph 12, when nucleotide sequences are circular in configuration, the applicant must choose the nucleotide in residue position number 1. For the purposes of this example, the “a” residue identified by the arrow in the figure will be used as position 1. However, any residue may be chosen as position 1. With the residue indicated by the arrow as position 1, the sequence should be included in a sequence listing as:

atggataatgaagaagttaacgaagaatgtatgagattatttttcaagaacgctcgtgcgcatctggataaacatctaacatcaaggttgacatgcgatgaaaatgaaaatgcatatatcacgttcagatgcttcctggatggaatacatcgcaaatctactaggtttctcgaagagctacttttgaaacaagaaaatatgtaccaccaatg (SEQ ID NO: 98)

The sequence should be further described using feature key “misc\_feature” with a location of “212^1”, which indicates that the last residue in the sequence, position 212, is linked to residue 1. A “note” qualifier must be included with a value indicating that the molecule is circular.

**Relevant ST.26 paragraphs:** 7(a), **12**, and Annex I, Section 5, Feature Key 5.15

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

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

A patent application describes the following compound:

segment A: ccugucgt-3’

segment B: uaguuguagaggccugucct-5’



wherein segment A and segment B are RNA sequences.

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

**YES – segment B**

**NO – segment A**

The enumerated sequence contains two segments of specifically defined nucleotides separated by the following “linker” structure:



The linker structure is not a nucleotide according to paragraph 3(g); therefore, each segment must be considered a separate sequence. Segment B contains more than 10 specifically defined nucleotides and ST.26 paragraph 7(a) requires inclusion in a sequence listing. Segment A contains only eight specifically defined nucleotides and therefore is not required to be included in a sequence listing.

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

**NO**

Segment A contains fewer than 10 specifically defined nucleotides, and as per ST.26 paragraph 8, it must not be included in a sequence listing.

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

Segment B is an RNA molecule; therefore, the element “INSDSeq\_moltype” must be “RNA.” The symbol “u” must not be used to represent uracil in an RNA molecule in a sequence listing. According to paragraph 14, the symbol “t” will be construed as uracil in RNA. Accordingly, segment B must be included in the sequence listing as:

tcctgtccggagatgttgat (SEQ ID NO: 40)

Thymine in RNA is considered a modified nucleotide, i.e. modified uracil, and must be represented in the sequence as “t” and be further described in a feature table. Accordingly, the thymine in position 1 must be further described using the feature key “modified\_base”, the qualifier “mod\_base” with “OTHER” as the qualifier value, and a qualifier “note” with “thymine” as the qualifier value.

The thymine, i.e. modified uracil, in position 1 should also be further described in a feature table using the feature key “misc\_feature” and a qualifier “note” with the value e.g., “The 5&apos; oxygen of the thymidine is attached through the linker (4-(3-hydroxybenzamido)butyl) phosphinic acid to another nucleotide sequence.” Where practicable, the other sequence may be directly indicated as the value in the qualifier “note”.

**Relevant ST.26 paragraphs:** 3(g), 7(a), 8, 13, **14**, 19, and 54

### *Paragraph 27 – The most restrictive ambiguity symbol should be used*

**Example 27-1: Shorthand formula for an amino acid**

(GGGz)2

Where z is any amino acid.

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

**YES**

The sequence is disclosed as a formula. (GGGz)2 is simply a shorthand way of representing the sequence GGGzGGGz. Conventionally, a sequence is expanded first, and the definition of any variable, i.e. “z”, is determined thereafter.

The sequence uses the nonconventional symbol “z”. The definition of “z” must be determined from the explanation of the sequence in the disclosure, which defines this symbol as any amino acid (see Introduction to this document). The example does not provide any constraint on “z”, e.g., that it is the same in each occurrence.

The peptide in the example has eight enumerated amino acids, six of which are specifically defined glycine residues, and the remaining two are the “z” variable that should be represented in this sequence using the conventional symbol “X”. ST.26 paragraph 7(b) requires inclusion of the sequence in a sequence listing as a single sequence with a single sequence identification number.

Note that the sequence is still encompassed by Paragraph 7(b) despite the fact that the enumerated and specifically defined residues are not contiguous.

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

The sequence uses the nonconventional symbol “z”, which according to the disclosure is any amino acid. The conventional symbol used to represent “any amino acid” is “X”. Therefore, the sequence must be represented as the single expanded sequence:

GGGXGGGX (SEQ ID NO: 41)

According to paragraph 27, “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. Since in this example “X” represents “any amino acid”, it must be annotated with the feature key VARIANT and a note qualifier with the value, “X can be any amino acid”.

Where practicable, each “X” should be annotated individually. However, a region of contiguous “X” residues, or a multitude of “X” residues dispersed throughout the sequence, may be jointly described with the feature key VARIANT using the syntax “x..y” as the location descriptor, where x and y are the positions of the first and last “X” residues, and a note qualifier with the value, “X can be any amino acid”.

Further, the example does not disclose that “z” is the same amino acid in both positions in the expanded sequence. However, if “z” is disclosed as the same amino acid in both positions, then a feature key “VARIANT” and a qualifier “note” should be provided stating that “X” in position 4 and 8 can be any amino acid, as long as they are the same in both positions.

**Relevant ST.26 paragraph(s):** 3(c), 7(b) and **27**

**Example 27-2: Shorthand formula - less than four specifically defined amino acids**

A peptide of the formula (Gly-Gly-Gly-z)n

The disclosure further states, that z is any amino acid and

(i) variable n is any length; or

(ii) variable n is 2-100, preferably 3

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

**NO**

Consideration of both disclosed embodiments (i) and (ii) of the enumerated peptide of the formula reveals that “n” can be “any length”; therefore, the most encompassing embodiment of “n” is indeterminate. Since “n” is indeterminate, the peptide of the formula cannot be expanded to a definite length, and therefore, the unexpanded formula must be considered.

The enumerated peptide in the unexpanded formula (“n” = 1) provides three specifically defined amino acids, each of which is Gly, and the symbol “z”. Conventionally “Z” is the symbol for “glutamine or glutamic acid”; however, the example defines “z” as “any amino acid” (see Introduction to this document). Under ST.26, an amino acid that is not specifically defined is represented by “X”. Based on this analysis, the enumerated peptide, i.e. GGGX, does not contain four specifically defined amino acids. Therefore, ST.26 paragraph 7(b) does not require inclusion, despite the fact that “n” is also defined as specific numerical values in some embodiments.

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

**YES**

The example provides a specific numerical value for variable “n,” i.e., a lower limit of 2, an upper limit of 100, and an exact value 3. Any sequence containing at least four specifically defined amino acids may be included in the sequence listing.

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

A sequence containing 100 copies of GGGX is preferred (SEQ ID NO: 42). A further annotation should indicate that up to 98 copies of GGGX could be deleted. Inclusion of further specific embodiments that are a key part of the invention is strongly encouraged.

According to paragraph 27, “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. Since “X” in SEQ ID NO: 42 represents “any amino acid”, it must be annotated with the feature key VARIANT and a note qualifier with the value, “X can be any amino acid”.

Where practicable, each “X” should be annotated individually. However, a region of contiguous “X” residues, or a multitude of “X” residues dispersed throughout the sequence, may be jointly described with the feature key VARIANT using the syntax “x..y” as the location descriptor, where x and y are the positions of the first and last “X” residues, and a note qualifier with the value, “X can be any amino acid”.

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

**Relevant ST.26 paragraph(s):** 3(c), 7(b), 26, and **27**

**Example 27-3: Shorthand formula - four or more specifically defined amino acids**

A peptide of the formula (Gly-Gly-Gly-z)n

Where z is any amino acid and variable n is 2-100, preferably 3.

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

**YES**

The enumerated peptide of the formula provides three specifically defined amino acids, each of which is Gly, and the symbol “z”. Conventionally, “Z” is the symbol for “glutamine or glutamic acid”; however, the description in this example defines “z” as “any amino acid” (see Introduction to this document). Under ST.26, an amino acid that is not specifically defined is represented by “X”. Based on this analysis, the enumerated repeat peptide does not contain four specifically defined amino acids. However, the description provides a specific numerical value for variable “n,” i.e., a lower limit of 2 and an upper limit of 100. Therefore, the example discloses a peptide having at least six specifically defined amino acids in the sequence GGGzGGGz, which is required by ST.26 to be included in a sequence listing.

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

Since “z” represents any amino acid, the conventional symbol used to represent the fourth and eighth amino acids is “X.”

ST.26 requires inclusion in a sequence listing of only the single sequence that has been enumerated by its residues. Therefore, at least one sequence containing any of 2, 3, or 100 copies of GGGX must be included in the sequence listing; however, the most encompassing sequence containing 100 copies of GGGX is preferred (SEQ ID NO: 42) (see Introduction to this document). In the latter case, a further annotation could indicate that up to 98 copies of GGGX could be deleted. Inclusion of two additional sequences containing 2 and 3 copies of GGGX, respectively (SEQ ID NO: 44-45), is strongly encouraged.

According to paragraph 27, “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. Since “X” in this example represents “any amino acid”, it must be annotated with the feature key VARIANT and a note qualifier with the value, “X can be any amino acid”.

Where practicable, each “X” should be annotated individually. However, a region of contiguous “X” residues, or a multitude of “X” residues dispersed throughout the sequence, may be jointly described with the feature key VARIANT using the syntax “x..y” as the location descriptor, where x and y are the positions of the first and last “X” residues, and a note qualifier with the value, “X can be any amino acid”.

Further, the example does not disclose that the “z” variable is the same in each of the two occurrences in the expanded sequence. However, if “z” is disclosed as the same amino acid in all locations, then a feature Key VARIANT and a Qualifier note should indicate that “X” in all positions can be any amino acid, as long as they are the same in all locations.

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

**Relevant ST.26 paragraph(s):** 3(c), 7(b), 26, and **27**

### *Paragraph 28 – Amino acid sequences separated by internal terminator symbols*

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

A patent application describes the following sequences:



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

**YES**

The application describes a nucleotide sequence, containing termination codons, which encodes three distinct amino acid sequences.

The enumerated nucleotide sequence contains more than 10 specifically defined nucleotides and must be included in a sequence listing as a single sequence.

Regarding the encoded amino acid sequences, paragraph 28 requires that amino acid sequences separated by an internal terminator symbol such as a blank space, must be included as separate sequences. Since each of “Protein A”, “Protein B”, and “Protein C” contain four or more specifically defined amino acids, ST.26 paragraph 7(b) requires that each must be included in a sequence listing and must be assigned its own sequence identification number.

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

The nucleotide sequence must be included in a sequence listing as:

caattcagggtggtgaatatggcgcccaatacgcaaaccgcctctccccgcgcgttggccgattcattaatgcagctggccaggcaggtgagcaggctggaaagcgggcagtgaatgaccatgattacggattcactggccgtcgttttacaacgtcgtgactgggaaaaccctggcgttacccaacttaatcgccttgcagcacattggtgtcaaaaataataataaccggatgtactatttatccctgatgctgcgtcgtcaggtgaatgaagtcgcttaagcaatcaatgtcggatgcggcgcgacgcttatccgaccaacatatcataa (SEQ ID NO: 46)

The nucleotide sequence should further be described using a “CDS” feature key for each of the three proteins and the element INSDFeature\_location must identify the location of each coding sequence, including the stop codon. In addition, for each “CDS” feature key, the “translation” qualifier should be included with the amino acid sequence of the protein as the qualifier value. The application does not disclose the genetic code table that applies to the translation (see Annex 1, Section 9, Table 7). If the Standard Code table applies, then the qualifier “transl\_table” is not necessary; however, if a different genetic code table applies, then the appropriate qualifier value from Table 7 must be indicated for the qualifier “transl\_table”. Finally, the qualifier “protein\_id” must be included with the qualifier value indicating the sequence identification number of each of the translated amino acid sequences.

The amino acid sequences must be included as separate sequences, each assigned its own sequence identification number:

MAPNTQTASPRALADSLMQLARQVSRLESGQ (SEQ ID NO: 47)

MTMITDSLAVVLQRRDWENPGVTQLNRLAAHWCQK (SEQ ID NO: 48)

MLRRQVNEVA (SEQ ID NO: 49)

NOTE: See “Example 90-1 Amino acid sequence encoded by a coding sequence with introns” for an illustration of a translated amino acid sequence represented as a single sequence.

**Relevant ST.26 paragraphs:** 7, 26, **28**, 57, 89-92

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

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

A patent application describes the following sequence:

Ala-Hse-X1-X2-X3-X4-Tyr-Leu-Gly-Ser

Wherein, X1= Ala or Gly,

X2= Ala or Gly,

X3= Ala or Gly,

X4= Ala or Gly, and

Hse = Homoserine

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

**YES**

The enumerated peptide contains five specifically defined amino acids. The symbol “X” is used conventionally to represent two amino acids in the alternative (see Introduction to this document).

Because there are five specifically defined amino acids, i.e., Ala, Tyr, Leu, Gly and Ser, ST.26 paragraph 7(b) requires that the sequence must be included in a sequence listing.

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

Paragraph 29 requires any “other” amino acid must be represented by the symbol “X”. In the example, the sequence contains the amino acid Hse in position 2 which is not found in Annex I, Section 3, Table 3. Accordingly, Hse is an “other” amino acid and must be represented by the symbol “X”.

X1-X4 are variant positions, each of which can be A or G. The most restrictive ambiguity symbol for alternatives A or G is “X”. Therefore, the sequence may be represented as:

AXXXXXYLGS (SEQ ID NO: 50)

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

Since amino acid Hse is not found in Annex I, Section 4, Table 4, a feature key “SITE” and a qualifier “note” must be provided with the complete, unabbreviated name of homoserine as per ST.26 paragraph 30.

According to paragraph 27, because X1-X4 represent an alternative of only 2 amino acids, then further description is required. Paragraph 96 indicates that the feature key “VARIANT” should be used with the qualifier “note” and qualifier value “A or G”. According to ST.26 paragraph 34, since these positions are adjacent and have the same description, they may be jointly described using the syntax “3..6” as the location descriptor in the element INSDFeature\_location.

**Relevant ST.26 paragraphs:** 3(a), 7(b), 25-27, **29**, 30, 34, 66, 70, 71, and 96-97

**Example 29-2: Use of the corresponding unmodified amino acid**

A patent application describes the following sequence:

Ala-Hyl-Tyr-Leu-Gly-Ser-Nle-Val-Ser-5ALA

Wherein Hyl = hydroxylysine (post-translational modification of lysine), Nle = Norleucine, and 5ALA = δ-Aminolevulinic acid

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

**YES**

The enumerated peptide contains more than four specifically defined amino acids; therefore, the sequence must be included in a sequence listing.

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

The hydroxylysine in position 2, the norleucine in position 7, and the δ-aminolevulinic acid in position 10 are all “modified amino acids”. First, we must consider each modified amino acid and determine if it should be represented by the corresponding unmodified amnio acid or the variable “X” in the sequence. Paragraph 29 states that a modified amino acid “should be represented in the sequence as the corresponding unmodified amino acids whenever possible.”

It is up to the discretion of the applicant to decide if a modified amino acid will be represented by the corresponding unmodified residue or the variables “X”. However, the following guidance should be taken into account: If an amino acid is modified by the addition of a moiety, such as methylation or acetylation, and the basic structure of the corresponding unmodified amino acid is generally unchanged, then representation by the unmodified amino acid is recommended. If the modified amino acid is structurally very different from the corresponding unmodified amino acid, then representation by an “X” is recommended.

The structure of hydroxylysine is nearly identical to lysine, except that the third carbon in the R-group is modified with a hydroxyl group. Since the basic structure of the corresponding unmodified lysine residue is intact, hydroxylysine should be represented in the sequence by lysine (“K”), not by “X”.

Norleucine is an isomer of leucine. The R-group of leucine is a 4 carbon chain, branched at the second carbon. Norleucine also has a 4 carbon R-group, but it is linear and not branched. Therefore, norleucine isn’t simply the result of a modification added to a leucine, but a completely different (although related) structure. Therefore, it is recommended that norleucine be represented by an “X” in a sequence listing.

δ-Aminolevulinic acid is not structurally similar to any of the amino acids listed in Annex I, Table 3. Therefore, it is recommended that δ-aminolevulinic acid be represented by an “X” in a sequence listing.

Accordingly, the sequence should be included in a sequence listing as:

AKYLGSXVSX (SEQ ID NO: 51)

Paragraph 30 requires the further annotation of each modified amino acid.

Hydroxylysine is a post-translational modification of lysine. Therefore, it must be described using the feature key “MOD\_RES” together with a qualifier “note” that describes the modification. Note that “hydroxylysine” is listed in Annex 1, Section 4, Table 4, “List of Modified Amino Acids.” Therefore, the value of the qualifier “note” can contain the abbreviation “Hyl” instead of the complete, unabbreviated name “hydroxylysine.”

Norleucine is not a post-translationally modified residue, therefore it must be described using the feature key “SITE” together with a qualifier “note” that descries the modification. Note that “norleucine” is also listed in Annex 1, Section 4, Table 4. Therefore, the value of the qualifier “note” can contain the abbreviation “Nle” instead of the complete, unabbreviated name “norleucine.”

δ-Aminolevulinic acid is also not a post-translationally modified residue, therefore it must be described using the feature key “SITE” together with a qualifier “note” that descries the modification. δ-Aminolevulinic acid is not listed in Annex 1, Section 4, Table 4, therefore, the value of the qualifier “note” must contain the complete, unabbreviated name of the modified residue, “δ-aminolevulinic acid.”

**Relevant ST.26 paragraphs**: 3(a), 3(e), 7(b), 29, and 30

### *Paragraph 30 – Annotation of a modified amino acid*

**Example 30-1 – Feature key “CARBOHYD”**

A patent application describes a polypeptide with a specifically modified amino acid, containing a glycosylated side chain, characterized in that Cys corresponding to positions 4 and 15 of the polypeptide forms a disulfide bond, according to the following sequence:

Leu-Glu-Tyr-Cys-Leu-Lys-Arg-Trp-Asn(asialyloligosaccharide)-Glu-Thr-Ile-Ser-His-Cys-Ala-Trp

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

**YES**

The enumerated peptide provides 17 specifically defined amino acids. There are 16 natural amino acids, wherein the ninth (asparagine) is glycosylated. Therefore, the sequence must be included in a sequence listing as required by ST.26 paragraph (7)(b).

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

According to ST.26 paragraph 29, a modified amino acid should be represented in the sequence as the corresponding unmodified amino acid whenever possible.

Therefore the sequence must be included in a sequence listing as:

LEYCLKRWNETISHCAW (SEQ ID NO: 52)

A further description of the modified amino acid is required. The feature key “CARBOHYD” together with the (mandatory) qualifier “note” should be used to indicate the occurrence of the attachment of a sugar chain (asialyloligosaccharide) to asparagine in position 9.  The qualifier “note” describes the type of linkage, e.g.,   
N-linked. The location descriptor in the feature location element is the residue position number of the modified asparagine.

In addition, there is a disulfide bond between the two Cys residues. Therefore the feature key “DISULFID” should be used to describe an intrachain crosslink. The feature location element is the residue position numbers of the linked Cys residues in “x..y” format, i.e., “4..15”. The mandatory qualifier note should describe the intrachain disulfide bond.

**Relevant ST.26 paragraph(s):** 3(a), 7(b), 26, 29, **30**, 66(c), 70, and Annex I, section 7, feature key 7.4

**Example 30-2 – Post-translationally modified amino acids**

A patent application describes the following polypeptide:

Leu-Glu-Tyr-Cys-Leu-Lys-Arg-Trp-Glu-Thr-Ile-Ser-His

wherein the Arg at position 7 may be post-translationally deiminated to citrulline.

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

**YES**

The enumerated peptide provides 13 specifically defined amino acids. Therefore, the sequence must be included in a sequence listing as required by ST.26 paragraph (7)(b).

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

According to ST.26 paragraph 29, a modified amino acid should be represented in the sequence as the corresponding unmodified amino acid whenever possible.

Therefore, the sequence should be included in a sequence listing as:

LEYCLKRWETISH (SEQ ID NO: 97)

where the symbol “R” is used to represent the arginine at position 7.

A further description indicating that the arginine at position 7 may be modified to citrulline is required. The modification of arginine to citrulline is a post-translational modification. Therefore, the feature key “MOD\_RES” should be used together with the mandatory qualifier “note” to indicate that the arginine may be deiminated to form citrulline. The location descriptor in the feature location element is the residue position number of the modified arginine.

**Relevant ST.26 paragraph(s):** 3(a), 7(b), **30**, and Annex I, Section 7, Feature Key 7.18

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

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

LL-100-KYMR

Where the “-100-“between amino acids Leucine and Lysine reflects a 100 amino acid region in the sequence.

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

**YES**

ST.26 paragraph 36 requires inclusion of a sequence that contains at least four specifically defined amino acids separated by one or more regions of a defined number of “X” residues.

The disclosed sequence uses a nonconventional symbol, i.e. “-100-.” The definition of “-100-” must be determined from the explanation of the sequence in the disclosure, which defines this symbol as 100 amino acids between leucine and lysine (see Introduction to this document). Therefore, “-100-” is a defined region of “X” residues. Since six of the 106 amino acids in the sequence are specifically defined, ST.26 paragraph 7(b) requires that the sequence must be included in a sequence listing.

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

The nonconventional symbol “-100-” is represented as 100 “X” residues (since any symbol used to represent an amino acid is equivalent to only one residue). Therefore, a single sequence of 106 amino acids in length, containing 100 “X” residues between LL and KYMR, must be included in a sequence listing (SEQ ID NO: 53).

This sequence contains 100 “X” variables between LL and KYMR. The ST.26 default value for “X” with no further annotation, is 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” (paragraph 27). If these 100 “X” variables are defined as anything other than this default value, then a proper annotation for each “X” variable must be provided.

**Relevant ST.26 paragraph(s):** 7(b), 26, 27, and **36**

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

Lys-z2-Lys-zm-Lys-z3-Lys-zn-Lys-z2-Lys

Where z is any amino acid, m=20, n=19-20, z2 means that the pairs of Lysines are separated by any two amino acids, and z3 means the pairs of Lysines are separated by any three amino acids.

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

**YES**

The disclosed sequence uses a nonconventional symbol, i.e. “z.” Therefore, the disclosure must be consulted to determine the definition; “z” is defined as any amino acid (see Introduction to this document). The conventional symbol used to represent any amino acid is “X”. Considering the presence of “X” variables, the peptide contains six lysine residues that are enumerated and specifically defined, which is required to be included in a sequence listing.

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

The sequence uses a nonconventional symbol “z”, the definition of which must be determined from the disclosure. Since “z” is defined as any amino acid, the conventional symbol is “X.”

The preferred and most encompassing means of representation is (see Introduction to this document):

KXXKXXXXXXXXXXXXXXXXXXXXKXXXKXXXXXXXXXXXXXXXXXXXXKXXK (SEQ ID NO: 54)

Wherein zn is equal to 20 “X’s”, with a further description that the “X” variable corresponding to position 30 can be deleted.

Alternatively, or in addition to the above, the sequence may be represented as:

KXXKXXXXXXXXXXXXXXXXXXXXKXXXKXXXXXXXXXXXXXXXXXXXKXXK (SEQ ID NO: 55)

Wherein zn is equal to 19 “X’s”, with a further description that an “X” variable between position numbers 29 and 30 can be inserted.

According to paragraph 27, “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. Since “X” in SEQ ID Nos 54 and 55 represents “any amino acid”, it must be annotated with the feature key VARIANT and a note qualifier with the value, “X can be any amino acid”.

Where practicable, each “X” should be annotated individually. However, a region of contiguous “X” residues, or a multitude of “X” residues dispersed throughout the sequence, may be jointly described with the feature key VARIANT using the syntax “x..y” as the location descriptor, where x and y are the positions of the first and last “X” residues, and a note qualifier with the value, “X can be any amino acid”.

**Relevant ST.26 paragraph(s):** 26, 27, and **36**

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

K-z2-K-zm-K-z3-K-zn-K-z2-K

Where z is any amino acid , where m=15-25, preferably 20-22, n=15-25, preferably 19-20, z2 means that the pairs of Lysines are separated by any two amino acids, and z3 means the pairs of Lysines are separated by any three amino acids.

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

**YES**

The sequence in the example uses a nonconventional symbol, i.e., “z.” Therefore, the surrounding disclosure is consulted to determine the definition of “z” (see Introduction to this document). The disclosure defines this symbol as any amino acid. The conventional symbol used to represent a residue defined as "any amino acid" is “X.” After considering the presence of “X” residues, the peptide contains 6 lysine residues that are enumerated and specifically defined, which is required in a sequence listing.

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

The sequence uses a nonconventional symbol “z”, the definition of which must be determined from the disclosure. Since “z” is defined as any amino acid, the conventional symbol is “X”. The preferred and most encompassing means of representation is:

KXXKXXXXXXXXXXXXXXXXXXXXXXXXXKXXXKXXXXXXXXXXXXXXXXXXXXXXXXXKXXK (SEQ ID NO: 56)

(where m=25 and n=25), with a further description that up to 10 “X” residues in each of the “zm” or “zn” regions may be deleted.

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

Alternatively, the sequence may be represented as:

KXXKXXXXXXXXXXXXXXXKXXXKXXXXXXXXXXXXXXXKXXK (SEQ ID NO: 57)

(where m=15 and n=15), with a further description that up to 10 “X” residues in each of the “zm” or “zn” regions may be inserted.

As further alternatives, any or all possible variations may be included.

According to paragraph 27, “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. Since “X” in SEQ ID Nos: 56 and 57 represents “any amino acid”, it must be annotated with the feature key VARIANT and a note qualifier with the value, “X can be any amino acid”.

Where practicable, each “X” should be annotated individually. However, a region of contiguous “X” residues, or a multitude of “X” residues dispersed throughout the sequence, may be jointly described with the feature key VARIANT using the syntax “x..y” as the location descriptor, where x and y are the positions of the first and last “X” residues, and a note qualifier with the value, “X can be any amino acid”.

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

**Relevant ST.26 paragraph(s):** 27 and **36**

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

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

Gly-Gly----Gly-Gly-Xaa-Xaa

where the symbol ---- is an undefined gap within the sequence, where Xaa is any amino acid, and the Glycine and Xaa residues are connected to one another through peptide bonds.

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

**NO**

ST.26 paragraph 37 prohibits the inclusion of any sequence that contains an undefined gap; therefore, inclusion of the entire sequence is not required.

ST.26 paragraph 37 does require inclusion of any region of a sequence adjacent to an undefined gap that contains four or more specifically defined amino acids. In the example above, inclusion of either region adjacent to the undefined gap is not required, since each region contains only two specifically defined amino acids.

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

**NO –** not the entire sequence

**NO –** not any region of the sequence

ST.26 paragraph 37 does not permit inclusion of the entire sequence.

ST.26 paragraph 8 does not permit inclusion of either region adjacent to the undefined gap, since each region contains only two specifically defined amino acids.

**Relevant ST.26 paragraphs:** 7(b), 8, 26, and **37**

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

Gly-Gly----Gly-Gly-Ala-Gly-Xaa-Xaa

wherein the symbol ---- is an undefined gap within the sequence, where Xaa is any amino acid, and the Glycine and Xaa residues are connected to one another through peptide bonds.

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

**NO –** not the entire sequence

**YES –** a region of the sequence

ST.26 paragraph 37 prohibits the inclusion of any sequence that contains an undefined gap, but requires inclusion of any region of a sequence adjacent to an undefined gap that contains four or more specifically defined amino acids.

In the example above, ST.26 does not require (and prohibits) inclusion of both the entire sequence, which contains an undefined gap, and the Gly-Gly region adjacent to the undefined gap, which contains only two specifically defined amino acids. However, ST.26 requires inclusion of the Gly-Gly-Ala-Gly- Xaa-Xaa region adjacent to the undefined gap, since it contains at least four specifically defined amino acids.

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

**NO –** not the entire sequence and not the Gly-Gly region

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

The region of the sequence adjacent to the undefined gap that contains four specifically defined amino acids must be represented as:

GGAGXX (SEQ ID NO: 58)

The sequence should be annotated to indicate that the represented sequence is part of a larger sequence that contains an undefined gap by using the feature key “SITE”, the feature location “1” and the qualifier “note” with the value, e.g., “This residue is linked N-terminally to a peptide having an N-terminal Gly-Gly and a gap of undefined length”.

According to paragraph 27, “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. Since “X” in SEQ ID NO: 58 represents “any amino acid”, it must be annotated with the feature key VARIANT and a note qualifier with the value, “X can be any amino acid”.

Where practicable, each “X” should be annotated individually. However, a region of contiguous “X” residues, or a multitude of “X” residues dispersed throughout the sequence, may be jointly described with the feature key VARIANT using the syntax “x..y” as the location descriptor, where x and y are the positions of the first and last “X” residues, and a note qualifier with the value, “X can be any amino acid”.

**Relevant ST.26 paragraph(s):** 7(b), 8, 26, 27, and **37**

*Paragraph 55 – A nucleotide sequence that contains both DNA and RNA segments*

**Example 55-1: Combined DNA/RNA Molecule**

A patent application describes the following oligonucleotide sequence:

AGACCTTcggagucuccuguugaacagauagucaaaguagauC

Wherein the upper-case letters represent DNA residues and lower-case letters represent RNA residues.

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

**YES**

The disclosed sequence has more than ten enumerated and specifically defined nucleotides; therefore, it is required to be included in a sequence listing.

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

The nucleotide sequence must be included in a sequence listing as:

agaccttcggagtctcctgttgaacagatagtcaaagtagatc (SEQ ID NO: 93)

Note that the uracil nucleotides must be represented by the symbol “t” in the sequence listing.

ST.26 paragraph 55 dictates that a nucleotide sequence containing both DNA and RNA segments must be indicated as molecule type “DNA” and must be further described 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”. In addition, each segment of the sequence must be further described with the feature key “misc\_feature,” which includes the location of the segment, and the qualifier “note,” which indicates whether the segment is DNA or RNA. The disclosed sequence contains two DNA segments (nucleotide positions 1-7 and 43) and one RNA segment (nucleotide positions 8-42).

**Relevant ST.26 paragraphs:** 7, 14, **55**-56, and 83

### *Paragraph 89 – “CDS” Feature key*

**Example 89-1: Encoding nucleotide sequence and encoded amino acid sequence**

A patent application describes the following nucleotide sequence and its translation:

atg acc gga aat aaa cct gaa acc gat gtt tac gaa att tta tga

Met Thr Gly Asn Lys Pro Glu Thr Asp Val Tyr Glu Ile Leu STOP

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

**YES**

The enumerated nucleotide sequence has more than ten specifically defined nucleotides.

The enumerated amino acid sequence has more than four specifically defined amino acids.

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

The nucleotide sequence must be presented as:

atgaccggaaataaacctgaaaccgatgtttacgaaattttatga (SEQ ID NO: 59)

The nucleotide sequence should further be described using the “CDS” feature key and the element INSDFeature\_location must identify the entire sequence, including the stop codon (i.e., position 1 through 45). In addition, the “translation” qualifier should be included with the qualifier value “MTGNKPETDVYEIL”. The application does not disclose the genetic code table that applies to the translation (see Annex 1, Section 9, Table 7). If the Standard Code table applies, then the qualifier “transl\_table” is not necessary; however, if a different genetic code table applies, then the appropriate qualifier value from Table 7 must be indicated for the qualifier “transl\_table”. Finally, the qualifier “protein\_id” must be included with the qualifier value indicating the sequence identification number of the translated amino acid sequence.

The amino acid sequence must be separately presented with its own sequence identification number using single letter codes as follows:

MTGNKPETDVYEIL (SEQ ID NO: 60)

The STOP following the enumerated amino acid sequence must not be included in the amino acid sequence in the sequence listing.

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

**Relevant ST.26 paragraphs:** 7(a), 7(b), 26, 28, 89, 90, and 92

**Example 89-2: Feature location extends beyond the disclosed sequence**

A patent application contains the following figure disclosing a partial coding sequence and its translated amino acid sequence:

cat cac gca gca gaa tgt gga ttt tgt cct caa caa tgg caa gtt cta 48

His His Ala Ala Glu Cys Gly Phe Cys Pro Gln Gln Trp Gln Val Leu

1 5 10 15

cgt ggg agt ctg tgc att tgt gag ggt cca gct gaa gga tgg ttc ata 96

Arg Gly Ser Leu Cys Ile Cys Glu Gly Pro Ala Glu Gly Trp Phe Ile

20 25 30

tca aga tgt tgg tta tgg tgt ggg cct caa gtc caa ggc ttt atc ttt 144

Ser Arg Cys Trp Leu Trp Cys Gly Pro Gln Val Gln Gly Phe Ile Phe

35 40 45

gga gaa ggc aag gaa gga ggc ggt gac aga cgg gct gaa gcg agc cct 192

Gly Glu Gly Lys Glu Gly Gly Gly Asp Arg Arg Ala Glu Ala Ser Pro

50 55 60

cag gag ttt tgg gaa tgc act tgg 216

Gln Glu Phe Trp Glu Cys Thr Trp

65 70

Figure 1 – partial coding sequence of the *Homo sapi*ens *ITCH1* gene, which encodes amino acids 20 through 91 of the 442 amino acid long ITCH1 protein.

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

**YES**

The application discloses a nucleotide sequence and its translated amino acid sequence. The enumerated nucleotide sequence contains more than 10 specifically defined nucleotides and must be included in a sequence listing.

The amino acid sequence contains more than 4 specifically defined amino acids and also must be included in a sequence listing as a separate sequence with its own sequence identification number.

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

The nucleotide sequence must be included in a sequence listing as:

catcacgcagcagaatgtggattttgtcctcaacaatggcaagttctacgtgggagtctgtgcatttgtgagggtccagctgaaggatggttcatatcaagatgttggttatggtgtgggcctcaagtccaaggctttatctttggagaaggcaaggaaggaggcggtgacagacgggctgaagcgagccctcaggagttttgggaatgcacttgg (SEQ ID NO: 94)

The nucleotide sequence should further be described using a “CDS” feature key. The element INSDFeature\_location must identify the location of the “CDS” feature in the sequence and must include the stop codon.

The figure describes a partial coding sequence that does not include the start codon or the stop codon. However, the description of the sequence indicates that the start codon is upstream of the nucleotide in position 1 and the stop codon is downstream of the last nucleotide in position 216.

ST.26 dictates that the location descriptor must not include numbering for residues beyond the range of the sequence in the INSDSeq\_sequence element. Consequently, in the above example, the location descriptor for the CDS feature key cannot include position numbers outside the range of 1 through 216. The location of the stop codon in the element INSDFeature\_location must be represented using the symbol “>” to indicate that the stop codon is located downstream of position 216. Likewise, the symbol “<” can be used to indicate that the location of the start codon is upstream of position 1. Thus, the location descriptor for the CDS feature key should appear as follows:

<1..>216

Note that “<” and “>” are reserved characters and will be replaced by “&lt;” and “&gt;”, respectively, in the XML instance of the sequence listing.

The “translation” qualifier should be included with the amino acid sequence of the protein as the qualifier value. The figure does not disclose the genetic code table that applies to the translation (see Annex 1, Section 9, Table 7). If the Standard Code table applies, then the qualifier “transl\_table” is not necessary; however, if a different genetic code table applies, then the appropriate qualifier value from Table 7 of ST.26 Annex I must be indicated for the qualifier “transl\_table”. Finally, the qualifier “protein\_id” must be included in the CDS feature with the qualifier value indicating the sequence identification number of the translated amino acid sequence.

The translated amino acid sequence must be included as a separate sequence with its own sequence identification number:

HHAAECGFCPQQWQVLRGSLCICEGPAEGWFISRCWLWCGPQVQGFIFGEGKEGGGDRRAEASPQEFWECTW (SEQ ID NO: 95)

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

**Relevant ST.26 paragraphs:** 7, 41, 65, 66, 70, 71, 89, and 92

### *Paragraph 92 – Amino acid sequence encoded by a coding sequence*

**Example 92-1: Amino acid sequence encoded by a coding sequence with introns**

A patent application contains the following figure disclosing a coding sequence and its translation:



Figure 1 – nucleotides shown in bold-face are intron regions.

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

**YES**

The application discloses a nucleotide sequence and its amino acid translation. The enumerated nucleotide sequence contains more than 10 specifically defined nucleotides and must be included in a sequence listing as a single sequence.

The nucleotide sequence contains coding sequence (exons) separated by noncoding sequence (introns). The figure depicts the translation of the nucleotide sequence as three non-contiguous amino acid sequences. According to the figure caption, the bolded regions of nucleotides are intron sequences that will be spliced out of an RNA transcript before translation into a protein. Accordingly, the three amino acid sequences are actually a single, contiguous, enumerated sequence, which contains more than four specifically defined amino acids and must be included in a sequence listing as a single sequence.

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

The nucleotide sequence must be included in a sequence listing as:

atgaagactttcgcagccttgctttccgctgtcactctcgcgctctcggtgcgcgcccaggcggctgtctggagtcaatgtaagtgccgctgcttttcattgatacgagactctacgccgagctgacgtgctaccgtataggtggcggtacaccgggttggacgggcgagaccacttgcgttgctggttcggtttgtacctccttgagctcagtgagcgactttcaatccgtcgtcattgctcctcatgtattgacgattggccttcatagtcatactctcaatgcgttccgggctccgcaacgtccagcgctccggcggccccctcagcgacaacttcaggccccgcacctacggacggaacgtgctcggccagcggggcatggccgccattgacctga (SEQ ID NO: 75)

The nucleotide sequence should further be described using a “CDS” feature key and the element INSDFeature\_location must identify the location of the coding sequence, including the stop codon indicated by “Ter”. The CDS INSDFeature\_location must use the “join” location operator to indicate that the translation products encoded by the indicated locations are joined and form a single, contiguous polypeptide using the format “join(x1..y1,x2..y2,x3..y3)”, e.g., “join(1..79,142..212,272..400)”. In addition, the “translation” qualifier should be included, with the amino acid sequence of the protein as the qualifier value. (Note that the terminator symbol “Ter” in the last position of the sequence must not be included in the amino acid sequence.) The application does not disclose the genetic code table that applies to the translation (see Annex 1, Section 9, Table 7). If the “Standard Code” table applies, then the qualifier “transl\_table” is not necessary; however, if a different genetic code table applies, then the appropriate qualifier value from Table 7 must be indicated for the qualifier “transl\_table”. Finally, the qualifier “protein\_id” must be included with the qualifier value indicating the sequence identification number of the translated amino acid sequence. The amino acid sequence must be included as a single sequence:

MKTFAALLSAVTLALSVRAQAAVWSQCGGTPGWTGETTCVAGSVCTSLSSSYSQCVPGSATSSAPAAPSATTSGPAPTDGTCSASGAWPPLT (SEQ ID NO: 76)

**Relevant ST.26 paragraphs:** 7, 26, 28, 57, 67, and 89-92

### *Paragraph 93 – Primary sequence and a variant, each enumerated by its residues*

**Example 93-1: Representation of enumerated variants**

The description includes the following sequence alignment.

*D. melanogaster* ACATTGAATCTCATACCACTTT

*D. virilis* ...-..G...C..--.G.....

*D. simulans* GT..G.CG..GT..SGT.G...

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

**YES**

It is common in the art to include “dots” in a sequence alignment to indicate “this position is the same as the position above it.” Therefore, the “dots” in *D.virilis* and *D. simulans* sequences are considered enumerated and specifically defined nucleotides, as they are simply a short-hand way of indicating that a given position is the same nucleotide as in *D. melanogaster*. In addition, sequence alignments frequently display the symbol “-“ to indicate the absence of a residue in order to maximize the alignment.

Accordingly, the nucleotide sequences of *D. melanogaster* and *D. simulans* contain twenty-two enumerated and specifically defined nucleotides, whereas the nucleotide sequence of *D. virilis* contains nineteen. Thus, each sequence is required by ST.26 paragraph 7(a) to be included in a sequence listing with separate sequence identification numbers.

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

*Drosophila melanogaster* sequence must be included in a sequence listing as:

acattgaatctcataccacttt (SEQ ID NO: 61)

*Drosophila virilis* sequence must be included in a sequence listing as:

acatggatcccacgacttt (SEQ ID NO: 62)

*Drosophila simulans* sequence must be included in a sequence listing as:

gtatggcgtcgtatsgtagttt (SEQ ID NO: 63)

**Relevant ST.26 paragraphs:** 7(a), 13, and 93

**Example 93-2: Representation of enumerated variants**

The description includes the following table of a peptide and functional variants thereof. A blank space in the table below indicates that an amino acid in the variant is the same as the corresponding amino acid in the “Sequence” and a “-“ indicates deletion of the corresponding amino acid in the “Sequence”.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Position | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Sequence | A | V | L | T | Y | L | R | G | E |
| Variant 1 |  |  |  |  |  |  |  |  | A |
| Variant 2 |  |  | P |  |  | P |  |  |  |
| Variant 3 |  |  | A | I | G | Y |  |  |  |
| Variant 4 |  |  |  |  |  |  | - |  |  |

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

**YES**

As indicated, a blank space in this table indicates that an amino acid in the variant is the same as the corresponding amino acid in the “Sequence”. Therefore, the amino acids of the variant sequences are enumerated and specifically defined.

Since the four variant sequences each contain more than four enumerated and specifically defined amino acids, each sequence is required by ST.26 paragraph 7(b) to be included in a sequence listing with separate sequence identification numbers.

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

AVLTYLRGE (SEQ ID NO: 77)

AVLTYLRGA (SEQ ID NO: 78)

AVPTYPRGE (SEQ ID NO: 79)

AVAIGYRGE (SEQ ID NO: 80)

AVLTYLGE (SEQ ID NO: 81)

**Relevant ST.26 paragraphs:** 7(b), 26, and 93

**Example 93-3: Representation of a consensus sequence**

A patent application includes Figure 1 with the following multiple sequence alignment.

*Consensus* LEG**n**EQFINA**ak**IIRHP**k**Y**nrk**T**ln**NDI**m**LIK

*Homo sapiens* LEG**N**EQFINA**AK**IIRHP**Q**Y**DRK**T**LN**NDI**M**LIK

*Pongo abelii* LEG**N**EQFINA**AK**IIRHP**Q**Y**DRK**T**VN**NDI**M**LIK

*Papio Anubis* LEG**T**EQFINA**AK**IIRHP**D**Y**DRK**T**LN**NDI**L**LIK

*Rhinopithecus roxellana* LEG**T**EQFINA**AK**IIRHP**N**Y**NRI**T**LD**NDI**L**LIK

*Pan paniscus* LEG**N**EQFINA**AK**IIRHP**K**Y**NRI**T**LN**NDI**M**LIK

*Rhinopithecus bieti* LEG**N**EQFINA**TK**IIRHP**K**Y**NGN**T**LN**NDI**M**LIK

*Rhinopithecus roxellana* LEG**N**EQFINA**TQ**IIRHP**K**Y**NGN**T**LN**NDI**M**LIK

The consensus sequence includes upper case letters to represent conserved amino acid residues, while the lower case letters “n”, “a”, “k”, “r”, “l” and “m” represent the predominant amino acid residues among the aligned sequences.

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

**YES**

The lower case letters in the consensus sequence each represent a single amino acid residue. Consequently, the consensus sequence, as well as each of the remaining seven sequences in Figure 1, includes at least four specifically defined amino acids. ST.26 paragraph 7(b) requires inclusion of all eight sequences in the sequence listing.

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

The lower case letters in the consensus sequence are being used as ambiguity symbols to represent the predominant amino acid among the possible variants for a specific position. Therefore, the lower case letters “n”, “a”, “k”, “r”, “l” and “m” are conventional symbols used in a nonconventional manner and the consensus sequence must be represented using an ambiguity symbol in place of each of the lower case letters.

The most restrictive ambiguity symbol should be used. For most positions in the consensus sequence, “X” is the most restrictive ambiguity symbol; however, the most restrictive ambiguity symbol for “D” or “N” in positions 20 and 25 is “B”. The consensus sequence should be included in the sequence listing as:

LEGXEQFINAXXIIRHPXYBXXTXBNDIXLIK (SEQ ID NO: 82)

According to paragraph 27, the symbol “X” will be construed as any one of “A”, “R”, “N”, “D”, “C”, “Q”, “E”, “G”, “H”, “I”, “L”, “K”, “M”, “F”, “P”, “O”, “S”, “U”, “T”, “W”, “Y”, or “V”, except where it is used with a further description in the feature table. Therefore, each “X” in the consensus sequence must be further described in a feature table using the feature key “VARIANT” and the qualifier “note” to indicate the possible variants for each position.

The remaining seven sequences must be included in the sequence listing as:

LEGNEQFINAAKIIRHPQYDRKTLNNDIMLIK (SEQ ID NO: 83)

LEGNEQFINAAKIIRHPQYDRKTVNNDIMLIK (SEQ ID NO: 84)

LEGTEQFINAAKIIRHPDYDRKTLNNDILLIK (SEQ ID NO: 85)

LEGTEQFINAAKIIRHPNYNRITLDNDILLIK (SEQ ID NO: 86)

LEGNEQFINAAKIIRHPKYNRITLNNDIMLIK (SEQ ID NO: 87)

LEGNEQFINATKIIRHPKYNGNTLNNDIMLIK (SEQ ID NO: 88)

LEGNEQFINATQIIRHPKYNGNTLNNDIMLIK (SEQ ID NO: 89)

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

**Relevant ST.26 paragraphs:** 7(b), 26, 27, 93, and 97

### *Paragraph 94 – Variant sequence disclosed as a single sequence with enumerated alternative residues*

**Example 94-1: Representation of single sequence with enumerated alternative amino acids**

A patent application claims a peptide of the sequence:

(i) Gly-Gly-Gly-[Leu or Ile]-Ala-Thr-[Ser or Thr]

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

**YES**

The sequence provides four specifically defined amino acids and ST.26 paragraph 7(b) requires inclusion of the sequence in a sequence listing.

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

Table 3 of Annex I, Section 3 defines the ambiguity symbol “J” as isoleucine or leucine. Therefore, the preferred representation of the sequence is:

GGGJATX (SEQ ID NO: 64)

which requires a further description in a feature table using the feature key “VARIANT” and the qualifier “note” to indicate that the “X” is serine or threonine.

Alternatively, the sequence may be represented, for example, as:

GGGLATS (SEQ ID NO: 65)

which requires a further description in a feature table using the feature key “VARIANT” and the qualifier “note” to indicate that L can be replaced by I, and S can be replaced by T.

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

**Relevant ST.26 paragraph(s):** 7(b), 8, 26, 27, 94, and 97

**Example 94-2 – Representation of single sequence with enumerated alternative amino acids that may be modified amino acids**

A patent application describes the following polypeptide:

Leu-Glu-Tyr-Cys-Leu-Lys-Arg-Trp-Xaa-Glu-Thr-Ile-Ser-His-Cys-Ala-Trp

where Xaa can be Ile, Ala, Phe, Tyr, aIle, MeIle, or Nle.

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

**YES**

The enumerated peptide provides 16 specifically defined amino acids. Therefore, the sequence must be included in a sequence listing as required by ST.26 paragraph (7)(b).

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

The most restrictive ambiguity symbol that can encompass “Ile, Ala, Phe, Tyr, aIle, MeIle, or Nle” is “X”. Therefore, the sequence must be included in a sequence listing as:

LEYCLKRWXETISHCAW (SEQ ID NO: 96)

ST.26 paragraph 30 requires that “[a] modified amino acid must be further described in the feature table”. However, paragraph 30 does not require any specific feature key be used to describe modified amino acids. While paragraph 30 describes the use of feature keys “CARBOHYD”, “LIPID”, “MOD\_RES”, and “SITE”, these feature keys are more appropriate for scenarios where the modified amino acid is not within a list of alternatives for a specific location. In this example, the feature key “VARIANT” satisfies the requirement of paragraph 30 since it allows for the inclusion of all of the alternatives for the variant site. So, the feature key “VARIANT” with the qualifier “note” and “Ile, Ala, Phe, Tyr, aIle, MeIle, or Nle” as a qualifier value should be used to describe the variant site at position 9. The use of a second feature key such as “SITE” with a qualifier “note” may be used to further identify the modified amino acids found at position 9.

**Relevant ST.26 paragraph(s):** 3(a), 7(b), 27, 30, **94**, 96, and Annex I, Section 4, Table 4

### *Paragraph 95(a) – A variant sequence disclosed only by reference to a primary sequence with multiple independent variations*

**Example 95(a)-1: Representation of a variant sequence by annotation of the primary sequence**

An application contains the following disclosure:

“Peptide fragment 1 is Gly-Leu-Pro-Xaa-Arg-Ile-Cys wherein Xaa can be any amino acid….

In another embodiment, peptide fragment 1 is Gly-Leu-Pro-Xaa-Arg-Ile-Cys wherein Xaa can be Val, Thr, or Asp….

In another embodiment, peptide fragment 1 is Gly-Leu-Pro-Xaa-Arg-Ile-Cys wherein Xaa can be Val.”

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

**YES**

“Peptide fragment 1” in each of the three disclosed embodiments provides at least six specifically defined amino acids; therefore, the sequence must be included in a sequence listing as required by ST.26 paragraph 7(b).

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

In this example, the enumerated sequence of “Peptide fragment 1” is disclosed three times, as three different embodiments, each with an alternative description of Xaa. In this example, “X” is the most restrictive ambiguity symbol for the Xaa position.

ST.26 requires inclusion of the disclosed enumerated sequence only once. In the most encompassing of the three embodiments, Xaa is any amino acid (see Introduction to this document). Therefore, the sequence that must be included in the sequence listing is:

GLPXRIC (SEQ ID NO: 66)

According to paragraph 27, “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. Since “X” in SEQ ID NO: 66 represents “any amino acid”, it must be annotated with the feature key VARIANT and a note qualifier with the value, “X can be any amino acid”.

Where practicable, each “X” should be annotated individually. However, a region of contiguous “X” residues, or a multitude of “X” residues dispersed throughout the sequence, may be jointly described with the feature key VARIANT using the syntax “x..y” as the location descriptor, where x and y are the positions of the first and last “X” residues, and a note qualifier with the value, “X can be any amino acid”.

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

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

GLPVRIC (SEQ ID NO: 67)

GLPTRIC (SEQ ID NO: 68)

GLPDRIC (SEQ ID NO: 69)

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

**Relevant ST.26 paragraph(s):** 7(b), 26, 27, and **95(a)**

### *Paragraph 95(b) – A variant sequence disclosed only by reference to a primary sequence with multiple interdependent variations*

**Example 95(b)-1: Representation of individual variant sequences with multiple interdependent variations**

A patent application describes the following consensus sequence:

cgaatg**n1**cccactacgaatg**n2**cacgaatg**n3**cccaca

wherein n1, n2, and n3 can be a, t, g, or c.

Several variant sequences are disclosed as follows:

if n1 is a, then n2 and n3 are t, g, or c;

if n1 is t, then n2 and n3 are a, g, or c;

if n1 is g, then n2 and n3 are t, a, or c;

if n1 is c, then n2 and n3 are t, g, or a.

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

**YES**

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

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

The enumerated sequence contains more than ten specifically defined nucleotides and three “n” residues. ST.26 requires inclusion of the disclosed enumerated sequence and where an ambiguity symbol is appropriate, the most restrictive symbol should be used. In this example, n1, n2, and n3 can be a, t, g, or c, so “n” is the most restrictive ambiguity symbol. Therefore, the sequence that must be included in the sequence listing is:

cgaatgncccactacgaatgncacgaatgncccaca (SEQ ID NO: 70)

ST.26 paragraph 15 states that “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 in the feature table. Since the value of every “n” residue in this sequence is equivalent to the default “a”, “c”, “g”, or “t”, no further annotation is required.

The enumerated sequence contains variations at three distinct locations and the occurrence of the variations is interdependent. Inclusion of additional sequences which represent additional embodiments that are a key part of the invention is **strongly** encouraged, as discussed in the introduction to this document. Therefore, according to ST.26 paragraph 95(b), the additional embodiments should be included in a sequence listing as four separate sequences, each with its own sequence identification number:

cgaatgacccactacgaatgbcacgaatgbcccaca (SEQ ID NO: 71)

cgaatgtcccactacgaatgvcacgaatgvcccaca (SEQ ID NO: 72)

cgaatggcccactacgaatghcacgaatghcccaca (SEQ ID NO: 73)

cgaatgccccactacgaatgdcacgaatgdcccaca (SEQ ID NO: 74)

(Note that b = t, g, or c; v = a, g, or c; h = t, a, or c; and d = t, g, or a; see Annex I, Section 1, Table 1)

According to ST.26 paragraph 15, the most restrictive symbol must be used to represent variable positions. Consequently, n2 and n3 must not be represented by “n” in the sequence.

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

**Relevant ST.26 paragraphs:** 7(a), 15, and 95(b)

[Appendix to Annex VI follows]

## **APPENDIX**

GUIDANCE DOCUMENT SEQUENCES IN XML

The Appendix is available at:

<https://www.wipo.int/standards/en/xml_material/st26/st26-annex-vi-appendix-guidance-document-sequences.xml>

[Annex VII follows]

# **ANNEX VII**

RECOMMENDATION FOR THE TRANSFORMATION OF A SEQUENCE LISTING FROM ST.25 TO ST.26:

POTENTIAL ADDED OR DELETED SUBJECT MATTER

*Version 1.6*

*Proposal presented for consideration by the Committee on WIPO Standards (CWS)*

*at its tenth session*

## Introduction

The requirements for the presentation of nucleotide and amino acid sequences differ between WIPO Standards ST.25 and ST.26. Consequently, the question has been raised as to whether Standard ST.26 would require addition or deletion of any subject matter in a sequence listing submitted as part of an international application under Standard ST.26 that may not be supported by an application from which priority is claimed.

## Scope of the Document

This document addresses the mandatory requirements of ST.26, and any potential consequences of those requirements. This document does not address every possible scenario; if the means of representation in ST.26, of information contained in an ST.25 sequence listing, is not clear, then the information may always be included in the application description to avoid deleted subject matter.

## Recommendations for Potential Added or Deleted Subject Matter

Review of the issues contained in this document demonstrates that transformation from ST.25 to ST.26 by itself should not inherently result in added or deleted subject matter, in particular, where the ST.25 sequence listing was fully compliant with Standard ST.25. However, there are certain scenarios that will require applicant caution. Recommendations have been provided to avoid added or deleted subject matter.

### *Scenario 1*

ST.25 uses numeric identifiers to tag various types of data, e.g., <110> for Applicant Name. ST.26 uses terms in the English language, as element names and attributes, for data tagging.

#### Recommendation:

The ST.26 terms simply describe the type of data content; therefore, the use of the ST.26 element names and attributes does not constitute added subject matter.

### *Scenario 2*

ST.26 explicitly requires inclusion of: (a) branched sequences; (b) sequences with D-amino acids; (c) nucleotide analogues; and (d) sequences with abasic sites. Under ST.25, the requirement for inclusion or the prohibition of such sequences is not clear.

#### Recommendation:

The disclosure contained in the application should be sufficient to represent these sequences in an ST.26 sequence listing, when they may not have been included in an ST.25 sequence listing. For certain types of information required by ST.26, care must be taken not to add subject matter beyond that disclosed, e.g., see discussion below (in Scenario 4) on the mol\_type qualifier for nucleotide sequences.

### *Scenario 3*

ST.26 excludes sequences with less than 10 specifically defined nucleotides (not including “n”) and less than 4 specifically defined amino acids (not including “X”).

#### Recommendation:

The excluded sequences may be included in the application body, where those sequences have not already been included therein.

### *Scenario 4*

For both nucleotide sequences and amino acid sequences, ST.26 has the mandatory feature key

"source" with two mandatory qualifiers.. ST.25 has a corresponding feature key for nucleotide sequences (which is rarely used) with no corresponding qualifiers and there is no corresponding feature key for amino acid sequences.

**Nucleotide sequences**

ST.26 - feature key 5.37 source; mandatory qualifier 6.39 mol\_type (see ST.26 paragraph 75)

|  |  |
| --- | --- |
| **Qualifier** | **Value** |
| mol\_type | genomic DNA |
| genomic RNA |
| mRNA |
| tRNA |
| rRNA |
| other DNA (applies to synthetic molecules) |
| other RNA (applies to synthetic molecules) |
| transcribed RNA |
| viral cRNA |
| unassigned DNA (applies where *in vivo* molecule is unknown) |
| unassigned RNA (applies where *in vivo* molecule is unknown) |

**Amino acid sequences**

ST.26 - feature key 7.30 source; mandatory qualifier 8.1 mol\_type (see ST.26 paragraph 75)

|  |  |
| --- | --- |
| **Qualifier** | **Value** |
| mol\_type | protein |

#### Recommendation:

The only issue of concern is the controlled vocabulary values associated with the mol\_type qualifier for nucleotide sequences. Some of the value choices listed above may not be sufficiently supported in the disclosure. Added subject matter may be avoided, however, by use of the most generic value for a particular sequence, e.g., “other DNA” and “other RNA” for a synthetic molecule and “unassigned DNA” and “unassigned RNA” for an *in vivo* molecule.

### *Scenario 5*

Where a sequence includes “Xaa”, ST.25 requires that further information concerning that residue be included in field <223>, which accompanies fields <221> (feature name) and <222> (feature location). ST.25 does not provide a default value for “Xaa” (“X” in ST.26). However, ST.26 does provide such a default value, and therefore, further information is not always required.  Two of the most frequently used annotations in peptide sequences is “any amino acid” or “any naturally occurring amino acid” for variable “Xaa” or “X”. This language could be interpreted to include amino acids other than those listed in the amino acid tables contained in either ST.25 or ST.26. The ST.26 default value for “X” with no further annotation, is any of the 22 individual amino acids listed in Annex I (see Section 3, Table 3).  This ST.26 default value may itself constitute added or deleted subject matter, and therefore, adversely affect the scope of a patent application when transitioning from ST.25 to ST.26.

#### Recommendations:

1. Where the ST.25 sequence listing includes a <221> feature name, <222> feature location corresponding to the Xaa, and <223> further information on Xaa, and the <221> feature name is also an appropriate ST.26 feature key, e.g., SITE, VARIANT, or UNSURE, then the ST.26 feature key should be used. Furthermore, to avoid potential deleted subject matter, the information in field <223> must be included in an accompanying qualifier “note”.
2. Where the ST.25 sequence listing includes a <221> feature name, <222> feature location corresponding to the Xaa, and <223> further information on Xaa, and the <221> feature name is not an ST.26 feature key, then ST.26 feature keys SITE or REGION, as appropriate, should be used. Furthermore, to avoid potential deleted subject matter, the information in field <223>, as well as the inappropriate <221> feature name, must be included in an accompanying qualifier “note”. For example, an ST.25 listing used a feature name that is not in ST.25 or ST.26, <221> Variable, together with further information <223> Xaa is any amino acid. In this example, the value of the ST.26 qualifier note would be “Variable – Xaa is any amino acid”.
3. Where the ST.25 sequence listing provides no <221>, <222>, or <223> field corresponding to the Xaa or where fields <221> and <222> corresponding to the Xaa are included, but no information is included in a corresponding <223> field (neither scenario is compliant with ST.25, but has occurred nonetheless), any information contained in the application body to describe “Xaa” should be included in the ST.26 qualifier “note” together with an appropriate feature key, e.g., SITE, REGION, or UNSURE, and location.

### *Scenario 6*

In ST.25, uracil is represented in the sequence by “u” and thymine is represented by “t”. In ST.26, uracil and thymine are both represented in the sequence by “t” and without further annotation; “t” represents uracil in RNA and thymine in DNA.

#### Recommendations:

1. Where a DNA sequence contains uracil, ST.26 considers it to be a modified nucleotide, and requires that uracil must be represented as a “t” and be further described using the feature key “modified\_base”, the qualifier “mod\_base” with “OTHER” as the qualifier value and the qualifier “note” with “uracil” as the qualifier value. This ST.26 annotation is not considered added subject matter where the ST.25 DNA sequence contained a “u”.
2. Where an RNA sequence contains thymine, ST.26 considers it to be a modified nucleotide, and requires that thymine must be represented as a “t” and be further described using the feature key “modified\_base”, the qualifier “mod\_base” with “OTHER” as the qualifier value and the qualifier “note” with “thymine” as the qualifier value. This ST.26 annotation is not considered added subject matter where the ST.25 RNA sequence contained a “t”.

### *Scenario 7*

In both ST.25 and ST.26, modified nucleotides or amino acids must have a further description. In ST.26, the identity of a modified nucleotide may be indicated using an abbreviation from Annex I, Section 2, Table 2, where applicable. Otherwise, the complete unabbreviated name of the modified nucleotide must be indicated. Similarly, the identity of a modified amino acid may be indicated using an abbreviation from Annex I, Section 4, Table 4, where applicable. Otherwise, the complete unabbreviated name of the modified amino acid must be indicated. In contrast, if a modified residue is not contained in an ST.25 table, use of the complete, unabbreviated name is not required, and not infrequently, an abbreviation is used instead.

#### Recommendations:

1. Where only an abbreviated name, which is not in Annex I, Section 2, Table 2 or Section 4, Table 4, was used both in the application and in an ST.25 sequence listing for either a modified nucleotide or a modified amino acid, and the abbreviated name is known in the art to reference only one specific modified nucleotide or modified amino acid, then use of the full, unabbreviated name would not itself constitute added subject matter.
2. Where only an abbreviated name, which is not in Annex I, Section 2, Table 2 or Section 4, Table 4, was used both in the application and in an ST.25 sequence listing for either a modified nucleotide or a modified amino acid (and the application contains no chemical structure), and the abbreviated name is not known in the art to reference one specific modified nucleotide or modified amino acid, i.e., the abbreviation is either not known at all in the art, or could possibly represent multiple different modified nucleotides or modified amino acids, then compliance with ST.26, without introduction of added subject matter, is not possible in this situation. Of course in this case, the priority application and sequence listing are themselves vague. To avoid potential deleted subject matter, the abbreviated name from the ST.25 sequence listing should be placed in an ST.26 “note” qualifier in addition to the value of the complete unabbreviated name of the modified nucleotide or modified amino acid. The complete unabbreviated name of the modified nucleotide or modified amino acid required in an ST.26 sequence listing will not be afforded priority to the earlier application. Care should be taken to draft the original (ST.25) sequence listing and application disclosure to include the unabbreviated name to avoid future issues.

### 

### *Scenario 8*

ST.25 contains a number of feature keys that are not contained in ST.26. Therefore, applicants must take care to capture the information contained in those ST.25 feature keys in a manner compliant with ST.26 without the introduction of added or deleted subject matter.

#### Recommendations:

The following table provides guidance as to the manner in which the information contained in a former ST.25 feature key may be included in compliance with ST.26 without the introduction of added or deleted subject matter. Numbers 1-23 are feature keys related to nucleotide sequences and numbers 24–43 are feature keys related to amino acid sequences.

| **No.** | **ST.25 Feature key <221>** | **ST.26 equivalent** | | |
| --- | --- | --- | --- | --- |
| **Feature key** | **Qualifier** | **Qualifier value** |
| 1 | allele | misc\_feature | allele | <223> value |
| 2 | attenuator | regulatory[[3]](#footnote-4) | regulatory\_class1 | “attenuator” |
| note *(if <223> present)* | <223> value |
| 3 | CAAT\_signal | regulatory1 | regulatory\_class1 | “CAAT\_signal” |
| note *(if <223> present)* | <223> value |
| 4 | conflict | misc\_feature | note | “conflict” and <223> value |
| 5 | enhancer | regulatory1 | regulatory\_class1 | “enhancer” |
| note *(if <223> present)* | <223> value |
| 6 | GC\_signal | regulatory1 | regulatory\_class1 | “GC\_signal” |
| note *(if <223> present)* | <223> value |
| 7 | LTR | mobile\_element1 | rpt\_type1 | “long\_terminal\_repeat” |
| note *(if <223> present)* | <223> value |
| 8 | misc\_signal | regulatory1 | regulatory\_class1 | “other” |
| note *(if <223> present)* | <223> value |
| 9 | mutation | variation | note | “mutation” and <223> value |
| 10 | old\_sequence | misc\_feature | note | “old\_sequence” and <223> value |
| 11 | polyA\_signal | regulatory1 | regulatory\_class1 | “polyA\_signal\_sequence” |
| note *(if <223> present)* | <223> value |
| 12 | promoter | regulatory1 | regulatory\_class1 | “promoter” |
| note *(if <223> present)* | <223> value |
| 13 | RBS | regulatory1 | regulatory\_class1 | “ribosome\_binding\_site” |
| note *(if <223> present)* | <223> value |
| 14 | repeat\_unit (a) when repeat\_region not used | misc\_feature | note | “repeat\_unit” and <223> value |
| repeat\_unit (b) when repeat\_region used | repeat\_region | rpt\_unit\_range | 1st residue..last residue |
| note *(if <223> present)* | <223> value |
| 15 | satellite | repeat\_region | satellite | “satellite” (or “microsatellite” or  “minisatellite” – if supported) |
| note *(if <223> present)* | <223> value |
| 16 | scRNA | ncRNA1 | ncRNA\_class1 | “scRNA” |
| note *(if <223> present)* | <223> value |
| 17 | snRNA | ncRNA1 | ncRNA\_class1 | “snRNA” |
| note *(if <223> present)* | <223> value |
| 18 | TATA\_signal | regulatory1 | regulatory\_class1 | “TATA\_box” |
| note | TATA\_signal and (if <223> present): <223> value |
| 19 | terminator | regulatory1 | regulatory\_class1 | “terminator” |
| note *(if <223> present)* | <223> value |
| 20 | 3’clip | misc\_feature | note | “3’clip” and <223> value |
| 21 | 5’clip | misc\_feature | note | “5’clip” and <223> value |
| 22 | -10\_signal | regulatory1 | regulatory\_class1 | “minus\_10\_signal” |
| note *(if <223> present)* | <223> value |
| 23 | -35\_signal | regulatory1 | regulatory\_class1 | “minus\_35\_signal” |
| note *(if <223> present)* | <223> value |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No.** | | **ST.25 Feature key <221>** | | **ST.26 equivalent** | | | | | | |
| **Feature key** | | | **Qualifier** | | | **Qualifier value** |
| 24 | | NON\_CONS | | This feature relates to a gap of an unknown number of residues in a single sequence, which is prohibited in both ST.25 (paragraph 22) and ST.26 (paragraph 37). Consequently, each region of specifically defined residues that is encompassed by ST.26 paragraph 7 must be included in the sequence listing as a separate sequence and assigned its own sequence identification number. To avoid added/deleted subject matter, each such sequence must be annotated to indicate that it is part of a larger sequence that contains an undefined gap. | | | | | | |
| REGION | | note | | Description | | |
| Description - as to where and to what the sequence is linked, e.g., this residue is linked N-terminally to a peptide having an N-terminal Gly-Gly and a gap of undefined length. | | | | | | |
| 25 | | SIMILAR | | REGION | | note | | “SIMILAR” and <223> value if  present | | |
| 26 | | THIOETH | | CROSSLNK | | note | | “THIOETH” and <223> value if present | | |
| For further location information guidance, see ST.26 Annex I, CROSSLNK Feature Key Comment | | | | | | |
| 27 | | THIOLEST | | CROSSLNK | | note | | “THIOLEST” and <223> value if  present | | |
| For further location information guidance, see ST.26 Annex I, CROSSLNK  Feature Key Comment | | | | | | |
| 28 | | VARSPLIC | | Discussed in a Scenario 13 below | | | | | | |
| 29 | | ACETYLATION | | MOD\_RES | | note | | “ACETYLATION” and <223> value if present | | |
| note | | Information required by ST.26 Annex I MOD\_RES Feature Key Comment, if possible (without added subject matter) | | |
| 30 | | AMIDATION | | MOD\_RES | | note | | “AMIDATION” and <223> value if present | | |
| note | | Information required by ST.26 Annex I MOD\_RES Feature Key Comment, if possible (without added subject matter) | | |
| 31 | | BLOCKED | | MOD\_RES | | note | | “BLOCKED” and <223> value if present | | |
| note | | Information required by ST.26 Annex I MOD\_RES Feature Key Comment, if possible (without added subject matter) | | |
| 32 | | FORMYLATION | | MOD\_RES | | note | | “FORMYLATION” and <223> value if present | | |
| note | | Information required by ST.26 Annex I MOD\_RES Feature Key Comment, if possible (without added subject matter) | | |
| 33 | | GAMMA-  CARBOXYGLUTAMIC  ACID  HYDROXYLATION | | MOD\_RES | | note | | “GAMMA-CARBOXYLGLUTAMIC ACID HYDROXYLATION” and <223> value if present | | |
| note | | Information required by ST.26 Annex I MOD\_RES Feature Key Comment, if possible (without added subject matter) | | |
| 34 | | METHYLATION | | MOD\_RES | | note | | “METHYLATION” and <223> value if present | | |
| note | | Information required by ST.26 Annex I MOD\_RES Feature Key Comment, if possible (without added subject matter) | | |
| 35 | | PHOSPHORYLATION | | MOD\_RES | note | | | “PHOSPHORYLATION” and <223> value if present | | |
| note | | | Information required by ST.26 Annex I MOD\_RES Feature Key Comment, if possible (without added subject matter) | | |
| 36 | | PYRROLIDONE  CARBOXYLIC ACID | | MOD\_RES | note | | | “PYRROLIDONE CARBOXYLIC ACID” and <223> value if present | | |
| note | | | Information required by ST.26 Annex I MOD\_RES Feature Key Comment, if possible (without added subject matter) | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **ST.25 Feature key <221>** | **ST.26 equivalent** | | |
| **Feature key** | **Qualifier** | **Qualifier value** |
| 37 | SULFATATION | MOD\_RES | note | “SULFATATION” and <223> value if present |
| note | Information required by ST.26 Annex I MOD\_RES Feature Key Comment, if possible (without added subject matter) |
| 38 | MYRISTATE | LIPID | note | “MYRISTATE” and <223> value if present |
| note | Information required by ST.26 Annex I LIPID Feature Key Comment, if possible (without added subject matter) |
| 39 | PALMITATE | LIPID | note | “PALMITATE” and <223> value if present |
| note | Information required by ST.26 Annex I LIPID Feature Key Comment, if possible (without added subject matter) |
| 40 | FARNESYL | LIPID | note | “FARNESYL” and <223> value if present |
| note | Information required by ST.26 Annex I LIPID Feature Key Comment, if possible (without added subject matter) |
| 41 | GERANYL-GERANYL | LIPID | note | “GERANYL-GERANYL” and <223> value if present |
| note | Information required by ST.26 Annex I LIPID Feature Key Comment, if possible (without added subject matter) |
| 42 | GPI-ANCHOR | LIPID | note | “GPI-ANCHOR” and <223> value if present |
| note | Information required by ST.26 Annex I LIPID Feature Key Comment, if possible (without added subject matter) |
| 43 | N-ACYL  DIGLYCERIDE | LIPID | note | “N-ACYL DIGLYCERIDE” and <223> value if present |
| note | Information required by ST.26 Annex I LIPID Feature Key Comment, if possible (without added subject matter) |

### *Scenario 9*

Certain feature keys present in both ST.25 and in ST.26, both for nucleotide sequences and amino acid sequences, have mandatory qualifiers in ST.26, as indicated below. The nucleotide sequence feature key “modified\_base” is also present in

both ST.25 and ST.26; however, Scenario 7 contains appropriate recommendations. ST.25 did not have any qualifiers, but did have a <223> free text field. When the information contained in an ST.25 <223> field is appropriate as the value for the ST.26 mandatory qualifier, then the information should be included as such. When an ST.25 <223> field has either not been provided or contains information that is not appropriate as the value for the ST.26 mandatory qualifier, then applicants must take care to capture the information contained in the ST.25 feature key/<223> field in a manner compliant with ST.26 without the introduction of added or deleted subject matter.

**Nucleotide sequences****[[4]](#footnote-5)**

|  |  |
| --- | --- |
| **Feature Key** | **Mandatory Qualifier** |
| 5.12 - misc\_binding | 6.3 - bound\_moiety |
| 5.30 - protein\_bind | 6.3 - bound\_moiety |

Recommendations:

1. If the ST.25 <223> field is absent or inappropriate, and the application description disclosed the name of the molecule/complex that may bind to the feature location of the nucleic acid, then that name should be included in the qualifier “bound\_moiety”.
2. Any information contained in the ST.25 <223> field that is inappropriate for inclusion in the qualifier “bound\_moiety” should be inserted into an appropriate optional qualifier of the feature key, e.g., “note”.
3. If the ST.25 <223> field is absent or inappropriate, and the application description did not disclose the name of the molecule/complex that may bind to the feature location of the nucleic acid, then the ST.26 feature key “misc\_feature” should be used instead of misc\_binding or protein\_bind, with the qualifier “note”.
4. If the ST.25 <223> field was absent, then the value of the qualifier “note” should be the name of the ST.25 feature key;
5. If the ST.25 <223> field contained inappropriate information, then the value of the qualifier “note” should be the name of the ST.25 feature key and the information from the <223> field.

**Amino acid sequences[[5]](#footnote-6)**

|  |  |
| --- | --- |
| **Feature Key** | **Mandatory Qualifier** |
| 7.2 – BINDING | 8.2 – note |
| 7.4 – CARBOHYD | 8.2 – note |
| 7.10 – DISULFID | 8.2 – note |
| 7.11 – DNA\_BIND | 8.2 – note |
| 7.12 – DOMAIN | 8.2 – note |
| 7.16 – LIPID | 8.2 – note |
| 7.17 – METAL | 8.2 – note |
| 7.18 – MOD\_RES | 8.2 – note |
| 7.23 – NP\_BIND | 8.2 – note |
| 7.29 – SITE | 8.2 – note |
| 7.39 – ZN\_FING | 8.2 – note |

#### Recommendations:

1. If the ST.25 <223> field is absent or inappropriate, and the application description disclosed the specific information required in the mandatory qualifier, then that information should be included in the mandatory qualifier “note”.
2. Any information contained in the ST.25 <223> field that is inappropriate for inclusion in the mandatory qualifier “note” (see feature key definition and comment) should be inserted into a second qualifier “note”.
3. If the ST.25 <223> field is absent or inappropriate, and the application description did not disclose the specific information required in the mandatory qualifier, then the ST.26 feature key “SITE” (for one amino acid) or “REGION” (for a range of amino acids) should be used instead, with the qualifier “note”.
4. If the ST.25 <223> field is absent, then the value of the qualifier “note” should be the name of the ST.25 feature key;
5. If the ST.25 <223> field contained inappropriate information, then the value of the qualifier “note” should be the name of the ST.25 feature key and the information from the <223> field.

### *Scenario 10*

Each specific feature key in ST.25 has a <222> field to indicate a feature location; however, ST.25 does not require an indication of the location for most features and the format of the location information is not standardized. Furthermore,

ST.25 does not have location operators, e.g., “join”. ST.26 has standardized location descriptors and operators and each feature must contain at least one location descriptor. (CDS features are a special case and are discussed below in Scenario 11).

#### Recommendations:

1. If the ST.25 sequence listing had a <222> field, direct importation or importation into ST.26 format should not raise any added subject matter consideration;
2. If the ST.25 sequence listing did not have a <222> field, but location information was contained in the application description, then direct importation or importation into ST.26 format should not raise any added subject matter consideration;
3. If neither the ST.25 sequence listing, nor the application description contained location information, then presumably, the feature applies to the entire sequence. (Indicating a location that is less than the entire sequence without support in the application description would likely constitute added/deleted subject matter.) Care should be taken to draft the original (ST.25) sequence listing and application disclosure to include location information to the extent possible to avoid future issues.

### *Scenario 11*

In ST.25, a coding sequence that encoded a single, contiguous polypeptide but that was interrupted by one or more non-coding sequence(s), e.g., introns, was indicated as multiple separate CDS features, as illustrated below:

<220>

<221> CDS

<222> (1)..(571)

<220>

<221> CDS

<222> (639)..(859)

In contrast, ST.26 has a join location operator that specifies that the polypeptides encoded by the indicated locations are joined and form a single, contiguous polypeptide. (Note: both ST.25 and ST.26 require that the stop codon be included in the CDS feature location.)

#### Recommendations:

1. If the ST.25 sequence listing or the application description clearly indicated that the polypeptide sequences encoded by the multiple separate CDS features form a single, contiguous polypeptide, then a coding sequence interrupted by an intron in a single CDS feature must be represented with the join location operator, as illustrated below, such that no added subject matter is introduced:

<INSDFeature\_key>CDS</INSDFeature\_key>

<INSDFeature\_location>join(1..571,639..859)</INSDFeature\_location>

1. If the ST.25 sequence listing or the application description did not indicate that the polypeptide sequences encoded by the two separate CDS features form a single, contiguous polypeptide, then use of the join location operator would likely constitute added subject matter.

### *Scenario 12*

ST.25 specifies that feature names must be one from Table 5 or 6. However, U.S. regulations indicated that these feature names were recommended, but not required. Therefore, a sequence in an ST.25 sequence listing (compliant with U.S. regulations) might have a “custom” feature key name with no corresponding feature key in ST.26. It is also possible that no feature name was provided for the <221> field or the <221> field is absent. These scenarios may be handled in a similar manner.

#### Recommendation:

The “custom” feature key name from ST.25 may be represented in an ST.26 sequence listing with no added subject matter as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type** | **ST.25 Feature Key <221>** | **Potential ST.26 Equivalent** | | |
| **Feature key** | **Qualifier** | **Qualifier value** |
| NA | “Custom” feature key | misc\_feature | note | “custom” feature key  name and <223> value if present |
| AA | “Custom” feature key | SITE or REGION | note | “custom” feature key  name and <223> value  if present |

### *Scenario 13*

ST.25 contains a feature key “VARSPLIC” defined as “description of sequence variants produced by alternative splicing”. In ST.26, “VARSPLIC” has been replaced with the broader feature key VAR\_SEQ defined as “description of sequence variants produced by alternative splicing, alternative promoter usage, alternative initiation and ribosomal frameshifting”. Therefore, the ST.26 sequence listing should not use “VAR\_SEQ” as a replacement of “VARSPLIC” without a further explanation.

#### Recommendation:

In ST.26 the feature “VAR\_SEQ” should be used with the qualifier “note”, whose value should include an explanation of the ST.25 narrower scope, e.g., “sequence variant produced by alternative splicing”. Any additional information contained in an accompanying ST.25 <223> field should also be included in the qualifier “note”.

### *Scenario 14*

If the source of a sequence was artificial, the ST.25 <213> Organism field requires the phrase “Artificial Sequence”. In ST.26, the feature key “source” requires the qualifier “organism”, whose value must be indicated as “synthetic construct”, rather than “Artificial Sequence”.

#### Recommendation:

The value for the ST.26 qualifier “organism” must be indicated as “synthetic construct”. To avoid potential deleted subject matter, any explanatory information contained in the required ST.25 <223> field should be included in a qualifier “note” (of the feature key “source”).

### *Scenario 15*

If the scientific name of the source organism of a sequence is unknown, the ST.25 <213> Organism field requires the term “Unknown”. In ST.26, the feature key “source” requires the qualifier “organism”, whose value must be indicated as “unidentified”, rather than “Unknown”.

#### Recommendation:

The value for the ST.26 qualifier “organism” must be indicated as “unidentified”. To avoid potential deleted subject matter, any explanatory information contained in the required ST.25 <223> field should be included in a qualifier “note” (of the feature key “source”).

### *Scenario 16*

ST.25 allows for the enumeration of amino acids to optionally include negative numbers, counting backwards starting with the amino acid next to number 1, for the amino acids preceding the mature protein, for example pre-sequences, pro-sequences, pre-pro-sequences and signal sequences. ST.26 does not allow for negative numbers in the feature location.

#### Recommendations:

1. If the ST.25 sequence listing had a feature or features represented in a <221> and an accompanying <222> field which contained negative and/or positive numbering, e.g., “PROPEP” and/or “CHAIN”, then in the ST.26 sequence listing, the appropriate feature key, e.g., “PROPEP” and/or “CHAIN”, should be used. A qualifier “note” may be used with the information in a <223> field, if any, as the qualifier value;
2. If the ST.25 sequence listing did not have a feature or features represented in a <221> and accompanying <222> field, but information was contained in the application description regarding the negative and/or positive numbering, then in the ST.26 sequence listing, the appropriate feature key, e.g., “PROPEP” and/or “CHAIN”, should be used. Otherwise, the feature key “REGION” may be used. A qualifier “note” may be used with information in the application description, if any, as the qualifier value;
3. If neither the ST.25 sequence listing, nor the application description, contains information explaining the negative and/or positive numbering, then to avoid potential deleted subject matter in the ST.26 sequence listing, the “REGION” feature key should be used, where the feature location spans the negatively numbered region of the ST.25 sequence. Also, a qualifier “note” should be used to indicate that the amino acid sequence was negatively numbered in the ST.25 sequence listing of the application to which priority is claimed.

### *Scenario 17*

ST.25 provides for publication information in fields <300> to <313>. ST.26 does not provide for inclusion of such information.

#### Recommendation:

The information contained in ST.25 fields <300> to <313> should be inserted into the accompanying application body, if not already contained therein.

### *Scenario 18*

ST.25 does not provide a standardized way to indicate that a CDS region of a nucleotide sequence was to be translated using a genetic code table other than the standard genetic code table. In contrast, ST.26 has a “transl\_table” qualifier that can be used with the “CDS” feature key to indicate that the region is to be translated using an alternative genetic code table. If the “transl\_table” qualifier is not used, the use of the standard genetic code table is assumed.

#### Recommendations:

1. If the ST.25 sequence listing or the application description clearly indicated that a CDS region is to be translated using an alternative genetic code table, then the “transl\_table” qualifier must be used with the appropriate genetic code table number as the qualifier value. Failure to use the “transl\_table” qualifier would likely constitute added subject matter, as the default “Standard Code” table would be assumed. Failure to include, in the ST.26 sequence listing, the alternative genetic code table information from the ST.25 sequence listing or from the application description would likely constitute deleted subject matter.
2. If the ST.25 sequence listing or the application description did not indicate that a CDS region is to be translated using an alternative genetic code table, then the “transl\_table” qualifier should not be used, or should be used only with the qualifier value “1,” i.e., the Standard Code table. Use of the “transl\_table” qualifier with any qualifier value other than “1” would likely constitute added and deleted subject matter.

### *Scenario 19*

ST.25 does not provide a standardized way to indicate the location of a feature, in particular, one contained in a site or region that extends beyond a specified residue or span of residues, e.g., a CDS region of a nucleotide sequence that extends beyond one or both ends of a disclosed sequence. In contrast, the ST.26 feature location descriptor provides a standardized way to indicate the location of such a site or region by using the “<“ or “>“ symbols. For example, the “CDS” feature location must include the stop codon, even when the stop codon is not included in the disclosed sequence itself, by indicating the location as e.g., 1..>321.

#### Recommendations:

1. Where the ST.25 sequence listing did not explicitly indicate that the location of a feature extended beyond the sequence, but such a location is either supported by the disclosure or is clear from the sequence itself, e.g., the stop codon of a CDS feature that is not contained in the sequence, then the “<“ or “>“ symbols may be used in the ST.26 sequence listing without addition of subject matter.
2. Where the ST.25 sequence listing did not explicitly indicate that the location of a feature extended beyond the sequence, and such a location is neither supported by the disclosure, nor is clear from the sequence itself, then compliance with ST.26, without introduction of added subject matter, may not be possible in this situation. In this case, the priority application and sequence listing are themselves arguably incomplete. In this situation, the location description of the feature in the ST.26 sequence listing will not be afforded priority to the earlier application. Care should be taken to draft the original (ST.25) sequence listing and application disclosure to include complete feature information.

### *Scenario 20*

ST.25 Appendix I requires that where a nucleotide sequence contains both DNA and RNA fragments, the value in <212> shall be “DNA” and the combined DNA/RNA molecule shall be further described in the <220> to <223> feature section; however, the exact nature of the further description is not clear and this requirement is not routinely followed. ST.26, paragraph 55, requires that each

DNA and RNA segment (ST.26 uses “segment” rather than “fragment” for internal consistency) of the combined DNA/RNA molecule must be further described with the feature key “misc\_feature”, which includes the location of the segment, and the qualifier “note”, which indicates whether the segment is DNA or RNA.

#### Recommendations:

1. If the ST.25 sequence listing described the DNA and RNA segments in one or more features using <221> misc\_feature, appropriate locations in <222>, and indications in <223> as to which segments were DNA or RNA, then incorporating that information into ST.26 format, using a misc\_feature for each DNA and RNA segment, should not raise any added subject matter consideration;
2. If the ST.25 sequence listing described the DNA and RNA segments in one or more features using a feature key in <221> other than misc\_feature, appropriate locations in <222>, and indications in <223> identifying which segments are DNA or RNA, then incorporating that information into ST.26 format, using a misc\_feature for each DNA and RNA segment and an additional “note” qualifier with the original <221> feature key as the value, should not raise any added or deleted subject matter consideration;
3. If the ST.25 sequence listing provides the identity (DNA or RNA) and location of each segment in a <223> field that is not associated with a <221> and <222> field, e.g., the explanation for an Artificial Sequence, then incorporating that information into ST.26 format using a misc\_feature for each DNA and RNA segment, should not raise any added subject matter consideration;
4. If the ST.25 sequence listing described the molecule in a feature using a <221> misc\_feature and a <223> noting that the molecule is a combined DNA/RNA molecule, but did not provide location information for each segment, and
   * 1. If the description provided the locations of each DNA and RNA segment, then incorporating that information into ST.26 format using a misc\_feature for each DNA and RNA segment, should not raise any added subject matter consideration;
     2. If the description does not contain the location information of each DNA and RNA segment, then compliance with ST.26, without introduction of added subject matter, may not be possible in this situation. In this case, the priority application and sequence listing are themselves arguably incomplete. In this situation, any location descriptions of the features in the ST.26 sequence listing will not be afforded priority to the earlier application. Care should be taken to draft the original (ST.25) sequence listing and application disclosure to include complete feature information.
5. If the ST.25 sequence listing described the molecule in a feature using a feature key in <221> other than misc\_feature and a <223> noting that the molecule is a combined DNA/RNA molecule, but did not provide location information for each segment, and
6. If the description provided the locations of each DNA and RNA segment, then incorporating that information into ST.26 format using a misc\_feature for each DNA and RNA segment and an additional “note” qualifier with the original <221> feature key as the value, should not raise any added or deleted subject matter consideration;
7. If the description does not contain the location information of each DNA and RNA segment, then compliance with ST.26, without introduction of added subject matter, may not be possible in this situation. In this case, the priority application and sequence listing are themselves arguably incomplete. In this situation, any location descriptions of the features in the ST.26 sequence listing will not be afforded priority to the earlier application. Care should be taken to draft the original (ST.25) sequence listing and application disclosure to include complete feature information.
8. If the ST.25 sequence listing noted that the molecule is a combined DNA/RNA molecule in a <223> field, e.g., the explanation for an Artificial Sequence, but did not provide any feature key or location information of each segment, and
9. If the description provided the locations of each DNA and RNA segment, then incorporating that information into ST.26 format using a misc\_feature for each DNA and RNA segment, should not raise any added subject matter consideration;
10. If the description does not contain the location information of each DNA and RNA segment, then compliance with ST.26, without introduction of added subject matter, may not be possible in this situation. In this case, the priority application and sequence listing are themselves arguably incomplete. In this situation, any location descriptions of the features in the ST.26 sequence listing will not be afforded priority to the earlier application. Care should be taken to draft the original (ST.25) sequence listing and application disclosure to include complete feature information.

[End of Annex VII and of Standard]

1. A numeric character reference refers to a character by its Universal Character Set/Unicode code point, and uses the format: “&#nnnn;” or “&#xhhhh;”, where “nnnn” is the code point in decimal form, and “hhhh” is the code point in hexadecimal form. [↑](#footnote-ref-2)
2. NOTE: While an application disclosure may represent nucleotides or amino acids with either lower case or upper case symbols, for a sequence included in a sequence listing, only lower case letters must be used for representation of a nucleotide sequence (see ST.26 paragraph 13) and only upper case letters must be used for representation of an amino acid sequence (see ST.26 paragraph 26). [↑](#footnote-ref-3)
3. 1 ST.26 may require that a specific ST.25 feature, e.g., TATA\_signal, be replaced by a broader feature key/qualifier/value, e.g., regulatory/regulatory\_class/TATA\_box. In such a case, in order to avoid addition of subject-matter that may lead to partial loss of priority, it is recommended to include the more limited term "TATA\_signal" in a "note" qualifier as shown in the above table (item N° 18). If in rare cases the Applicant considers that the use of the "TATA\_box" value for the "regulatory\_class" qualifier is not appropriate, the value:"other" may be used instead of "TATA\_box". In this case, the term "TATA\_signal" must be included in a "note" qualifier associated to the “regulatory" feature key.. [↑](#footnote-ref-4)
4. The numeric references in the table below refer to the Feature key and Qualifier numbers of ST.26, Annex I Controlled Vocabulary. [↑](#footnote-ref-5)
5. The numeric references in the table below refer to the Feature key and Qualifier numbers of ST.26, Annex I Controlled Vocabulary. [↑](#footnote-ref-6)