Draft ST.26 Guidance Document

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.

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.

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.

Preparation of a sequence listing

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

1. Does ST.26 paragraph 6 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 6 (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 7 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(b*bis*) 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 and the amino acid symbols set forth in Section 3, Table 3 of ST.26 Annex 1. Symbols other than those set forth in these tables are "**nonconventional"**.

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

Where a conventional symbol or abbreviation is used, the explanation of the sequence in the disclosure must still be consulted to confirm that the symbol is used in a conventional manner. Otherwise, if the symbol is used in a nonconventional manner, the explanation is necessary to determine whether ST.26 paragraph 6 requires inclusion in the sequence listing or whether paragraph 7 prohibits inclusion.

Where a nonconventional symbol or abbreviation is disclosed as equivalent to a conventional symbol or abbreviation (e.g., "Z_1" means "A"), or to a specific sequence of conventional symbols (e.g., "Z_1" 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 6 requires inclusion in the sequence listing or whether paragraph 7 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 6 requires inclusion of the sequence in the sequence listing or whether ST.26 Paragraph 7 prohibits inclusion. Similarly, where a nonconventional amino acid symbol is used as an ambiguity symbol (e.g., "Z_1" 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 6 requires inclusion of the sequence in the sequence listing or whether ST.26 paragraph 7 prohibits inclusion.

"Specifically defined"

ST.26 paragraph 3(h) 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 6 is disclosed by enumeration of its residues only once in an application, but is described differently in multiple embodiments, e.g. in one embodiment "X" in one or more locations could be any amino acid, but in further embodiments, "X" could be only a limited number of amino acids, ST.26 requires inclusion in a sequence listing of only the single sequence that has been enumerated by its residues. As per paragraphs 15 and 26, where such a sequence contains multiple "n" or "X" ambiguity symbols, "n" or "X" is construed to represent any nucleotide or amino acid, respectively, in the absence of further

annotation. Consequently, the single sequence required to be included is the most encompassing sequence disclosed. The **most encompassing sequence** is the single sequence having variant residues which are represented by the most restrictive ambiguity symbols that include the most disclosed embodiments. However, inclusion of additional specific sequences is *strongly* encouraged where practical, e.g. which represent additional embodiments that are a key part of the invention. Inclusion of the additional sequences allows for a more thorough search and provides public notice of the subject matter for which a patent is sought.

Proper Usage of the Ambiguity Symbol "n" in a Sequence Listing

The symbol "n"

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

Proper Usage of the Ambiguity Symbol "X" in a Sequence Listing

The symbol "X"

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

ii. "unknown" amino acid, i.e., not determined, not disclosed, or unsure; or

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

Table A – Conventional Nucleotide Symbols, Abbreviations, and Names

Symbol	Abbreviation	Nucleotide Name
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 B – Conventional Amino Acid Symbols, Abbreviations, and Names

Symbol	3-Letter Abbreviation	Amino Acid Name
A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic Acid (Aspartate)
C	Cys	Cysteine
E	Glu	Glutamic Acid (Glutamate)
Q	Gln	Glutamine
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”

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 (6)(b).

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

Paragraph 28 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 24 indicates that when amino acid sequences are circular in configuration, applicant must choose the amino acid in residue position number 1. Accordingly, the sequence can be represented as:

AEKXGMX (SEQ ID NO: 1)

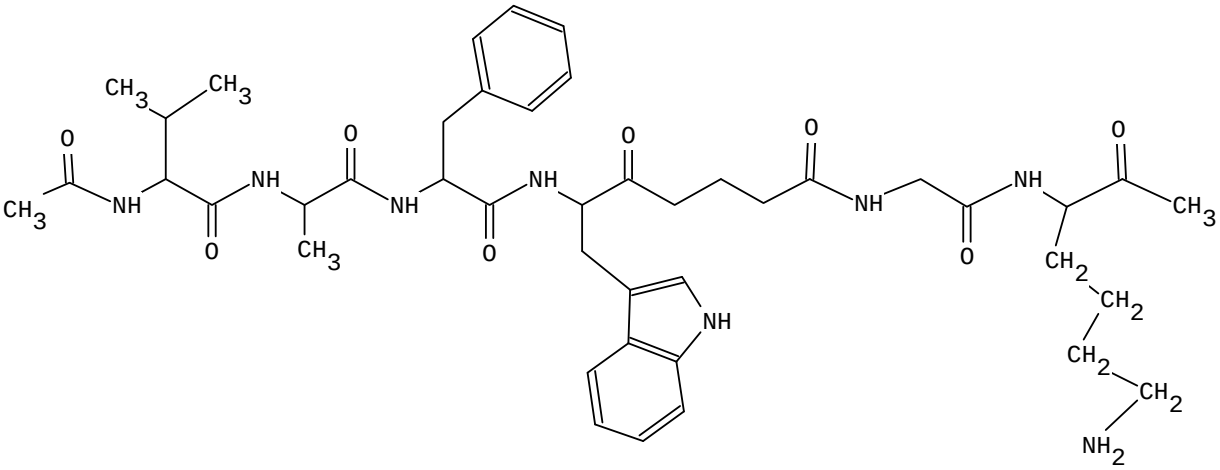
or otherwise, with any other amino acid in the sequence in residue position number 1. A feature key "SITE" and a qualifier "NOTE" must be provided for each D-amino acid with the complete, unabbreviated name of the D-amino acid as the qualifier value, e.g., D-Alanine and D-Norleucine. Further, a feature key

"SITE" and a qualifier "NOTE" must be provided with the abbreviation for L-norleucine as the qualifier value, i.e. "Nle", as set forth in Annex I, Section 4, Table 4. Finally, a feature key "REGION" and a qualifier "NOTE" should be provided to indicate that the peptide is circular.

Relevant ST.26 paragraphs: Paragraphs **3(a)**, 6(b), 24, 25, 28, 29, and 30

Paragraph 3(b*bis*) – Definition of "enumeration of its residues"

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



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

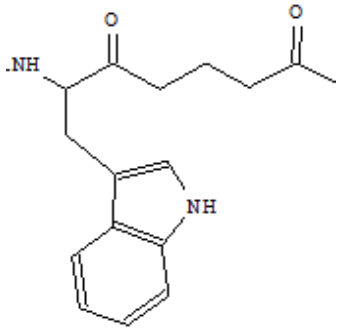
YES

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 can be represented as:

VAFXGK (SEQ ID NO: 2)



wherein "X" represents an "other" modified amino acid: , which requires a feature key "SITE" together with the qualifier "NOTE". The qualifier "NOTE" provides the complete, unabbreviated name of the modified tryptophan in

position 4 of the enumerated peptide, e.g., “6-amino-7-(1H-indol-3-yl)-5-oxoheptanoic acid”. Further, additional feature keys “SITE” and qualifier “NOTE” are required to indicate the acetylation of the N-terminus and the methylation of the C-terminus.

Alternatively, the sequence can be represented as:

VAFW (SEQ ID NO: 3)

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

Relevant ST.26 paragraph(s):

Paragraphs **3(b*bis*)**, 6(b), 28, 29, and 30

Example 3(b*bis*)-2: Shorthand formula for an amino acid sequence

$(G_4z)_n$

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

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

Yes

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

The enumerated peptide in the unexpanded formula ("n" = 1) provides four specifically defined amino acids, each of which is Gly, and the symbol "z". Conventionally "Z" is the symbol for "glutamine or glutamic acid"; however, the example defines "z" as "any amino acid". Under ST.26, an amino acid that is not specifically defined is represented by "X". Based on this analysis, the enumerated peptide, i.e. GGGGX, contains four glycine residues that are enumerated and specifically defined. Thus, ST.26 paragraph 6(b) requires inclusion of the sequence in a sequence listing.

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

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

GGGGX (SEQ ID NO: 4)

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

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

Relevant ST.26 paragraph(s): Paragraph **3(b*bis*)** and 6(b)

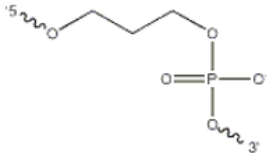
Paragraph 3(d) Definition of "nucleotide"

Example 3(d)-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(d); 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 SEQ ID 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 sequence.

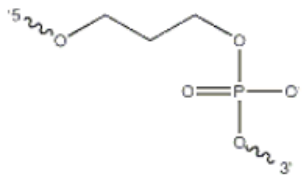
Relevant ST.26 paragraphs: Paragraphs **3(d)**, 6(a), and 15

Example 3(d)-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, ST.26 also requires inclusion of the two 12 nucleotide sequences.

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” may 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 highly recommended, 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 SEQ ID 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 sequence and identify the spacer by either its complete unabbreviated chemical name, or by its common name, e.g. C3 spacer.

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

Relevant ST.26 paragraphs: Paragraphs **3(d)**, 6(a), and 15

Example 3(d)-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(d)(ii)B 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 (6)(b).

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

The sequence must be included in a sequence listing as:

gagcattgacntaaggct (SEQ ID NO: 10)

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

Relevant ST.26 paragraphs: Paragraphs **3(d)**, 6(a), and 17

Example 3(d)-4: Nucleic Acid Analogues

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

PO_4-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(d)(i)(B). Accordingly, the sequence has more than ten enumerated and "specifically defined" nucleotides and is required to be included in a sequence listing.

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

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

tagttcattgactaaggctccccattgact (SEQ ID NO: 11)

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

Relevant ST.26 paragraphs: Paragraphs 3(cter), **3(d)**, 6(a), 10, 16, 17bis, 66, and 67

Paragraph 3(h) Definition of "specifically defined"

Example 3(h)-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(h), 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 6(a) to be included in a sequence listing.

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

NO

According to paragraph 7, "A sequence listing must not include any sequences having fewer than ten specifically defined nucleotides…." The enumerated sequence does not have ten or more specifically defined nucleotides; therefore, it must not be included in a sequence listing.

Relevant ST.26 paragraphs: Paragraphs **3(h)**, 6(a), 7, and 13

Example 3(h)-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 6(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 SEQ ID 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, at location 10 the feature key “misc_difference” should be used together with two “replace” qualifiers where the value for one would be “g” and the second would be “c”.

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

Relevant ST.26 paragraphs: Paragraphs **3(h)**, 6(a), 7, and 13

Example 3(h)-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(h), 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'-aatgttgga**c**-3'

This sequence has ten specifically defined nucleotides and is required by ST.26 paragraph 6(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: Paragraphs **3(h)** and 6(a)

Example 3(h)-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(h), 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 6(a) to be included in a sequence listing.

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

The sequence must be included in a sequence listing as:

nngkngkngkagvcr (SEQ ID NO: 16)

Relevant ST.26 paragraphs: Paragraphs **3(h)**, 6(a) and 15

Example 3(h)-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 6(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:

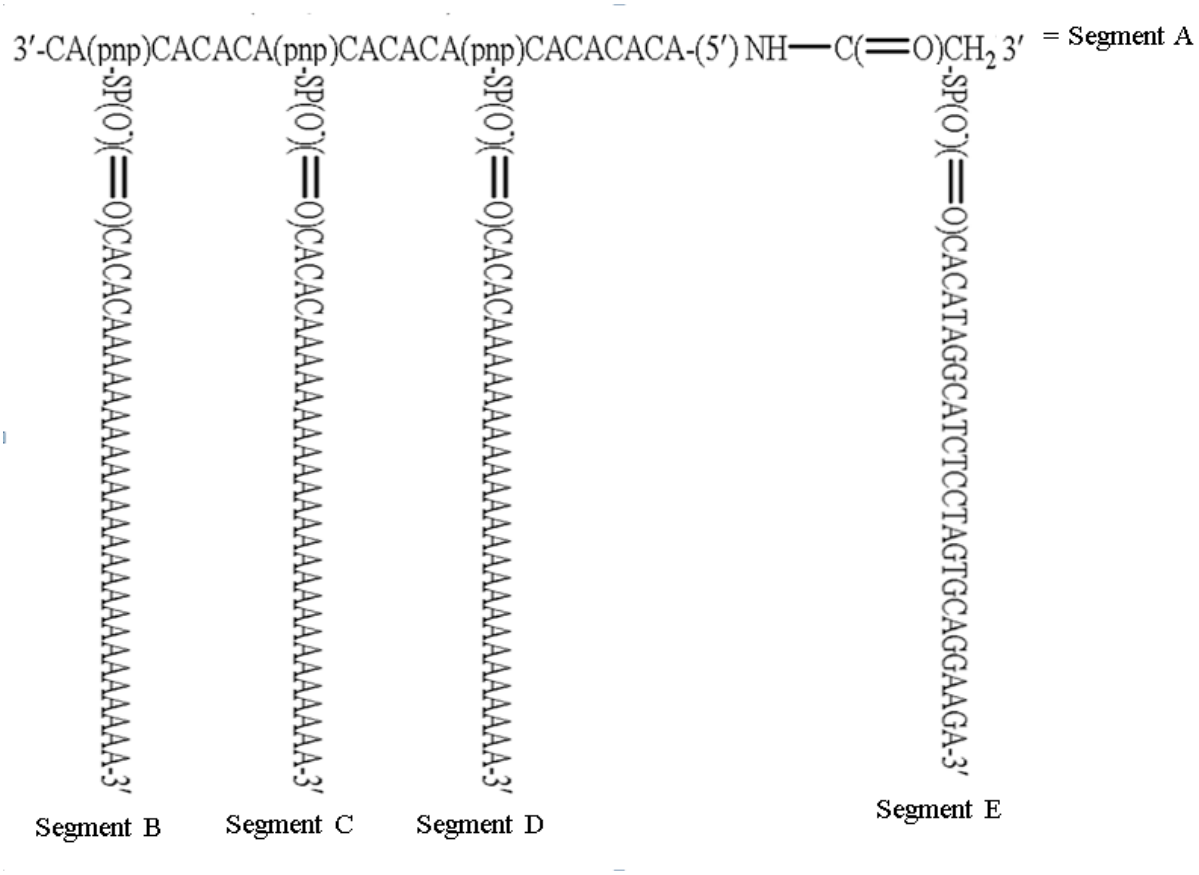
XYEKGJL (SEQ ID NO: 17)

Relevant ST.26 paragraphs: Paragraphs **3(h)**, 6(b), 25, and 26

Paragraph 6(a) – Nucleotide sequences required in a sequence listing

Example 6(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)CH_2 3’ is segment A;

SP(O^-)(=O)CACACAAAAAAAAAAAAAAAAAAAA 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 6(a), the linear portions of branched nucleotide sequences containing ten or more specifically defined nucleotides, wherein adjacent nucleotides are joined 3' to 5', must be included in a sequence listing.

The four vertical segments B-E each contain more than ten specifically defined nucleotides, wherein adjacent nucleotides are joined 3' to 5', and therefore each is required to be included in a sequence listing.

In horizontal segment A, the linear portions of the nucleotide sequence are linked by the non-nucleotide moiety "pnp" and each of these linked linear portions contains fewer than ten specifically defined nucleotides. Therefore, since no portion of segment A contains ten or more specifically defined nucleotides wherein adjacent nucleotides are joined 3' to 5', they are not required ST.26 paragraph 6(a) to be included in a sequence listing.

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

According to paragraph 7, "A sequence listing must not include any sequences having fewer than ten specifically defined nucleotides…."

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

However, segments B, C, D, and E can 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:

cacacaaaaaaaaaaaaaaaaaaaaaaaa. (SEQ ID NO: 18)

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

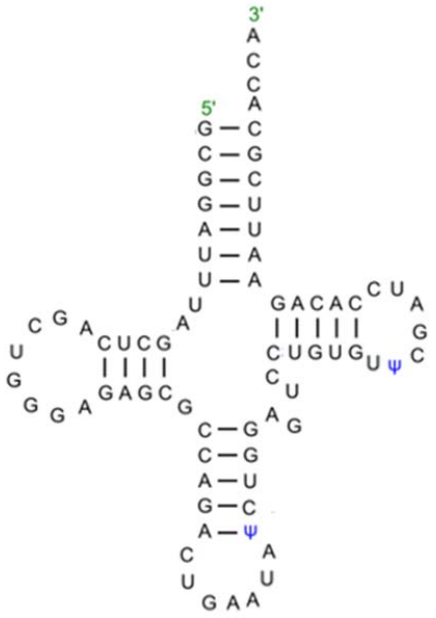
Segment E must be included in a sequence listing as a single sequence:

cacataggcatctcctagtgcaggaaga. (SEQ ID NO: 19)

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

Relevant ST.26 paragraph(s): Paragraphs **6(a)**, 7, 10, 13, and 17

Example 6(a)-2: Linear nucleotide sequence having a secondary structure



Wherein Ψ is pseudouridine.

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

YES

The nucleotide sequence contains seventy-three enumerated and specifically defined nucleotides. Thus, the example has ten or more "specifically defined" nucleotides, and as required by ST.26 paragraph (6)(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" cannot 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:

gcggatttagctcagctgggagagcgccagactgaatanctggagtcctgtgtncgatccacagaattcgcacc
a (SEQ ID NO: 20)

The value of the mandatory “mol_type” qualifier of the mandatory “source” feature key is “tRNA”. Additional information can be provided with feature key “tRNA” and any appropriate qualifier(s).

The “n” residues must be further described in a feature table using the feature key “modified_base” and the mandatory qualifier “mod_base” with the abbreviation “p” for pseudouridine as the qualifier value (see Annex 1, Table 2).

Relevant ST.26 paragraph(s): Paragraphs **6(a)**, 10, 13, 14, 62, 85 and Annex I, sections 2 and 5, feature key 5.42

Example 6(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 6(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:

gatcatttttttatatttttttatatttttttatatttttttatgtac (SEQ ID NO: 21)

Relevant ST.26 paragraphs: Paragraph **6(a) and 13**

Example 6(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 6(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: Paragraph **6(a) and 13**

Example 6(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 6(a) to be included in a sequence listing.

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

The sequence must be included in a sequence listing as:

gatcatacgcactatacgcactatacgcactatacgcactgtac (SEQ ID NO: 23)

Relevant ST.26 paragraphs: Paragraph **6(a) and 13**

Example 6(a)-6: Nonconventional nucleotide symbols

A patent application describes the following sequence:

5’ GATC-β-β-β-β-GTAC 3’

The explanation of the sequence in the disclosure further indicates that “β” is equal to adenine, inosine, or pseudouridine.

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

NO

The enumerated sequence uses the nonconventional symbol “β”. The explanation of the sequence in the disclosure must be consulted to determine the meaning of “β” (*see* Introduction to this document).

Consultation of the disclosure indicates that “β” is equivalent to adenine, inosine, or pseudouridine. The only conventional symbol that can be used to represent “adenine, inosine, or pseudouridine” is “n”; therefore, the “β” is a nonconventional symbol used to represent the conventional symbol “n”. Accordingly, the sequence must be interpreted to have four “n” symbols in place of the four “β” symbols:

5’ GATC-N-N-N-N-GTAC 3’

The enumerated sequence has only eight specifically defined nucleotides and is not required by ST.26 paragraph 6(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’ cannot be included in a sequence listing.

However, a disclosed alternative sequence could 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” are also needed. The value for the “mod_base” qualifier can be “OTHER” with a “note” qualifier and the value of “i or p”.

Other permutations are possible.

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

Relevant ST.26 paragraphs: Paragraph **6(a)**, 7, and 13

Paragraph 6(b) – Amino Acid sequences required in a sequence listing

Example 6(b)-1: Four or more specifically defined amino acids

XXXXXXXXDXXXXXXXXXXFXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXAXXXXXXXXXXXXXXXXXXGXXXXX

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 6(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:

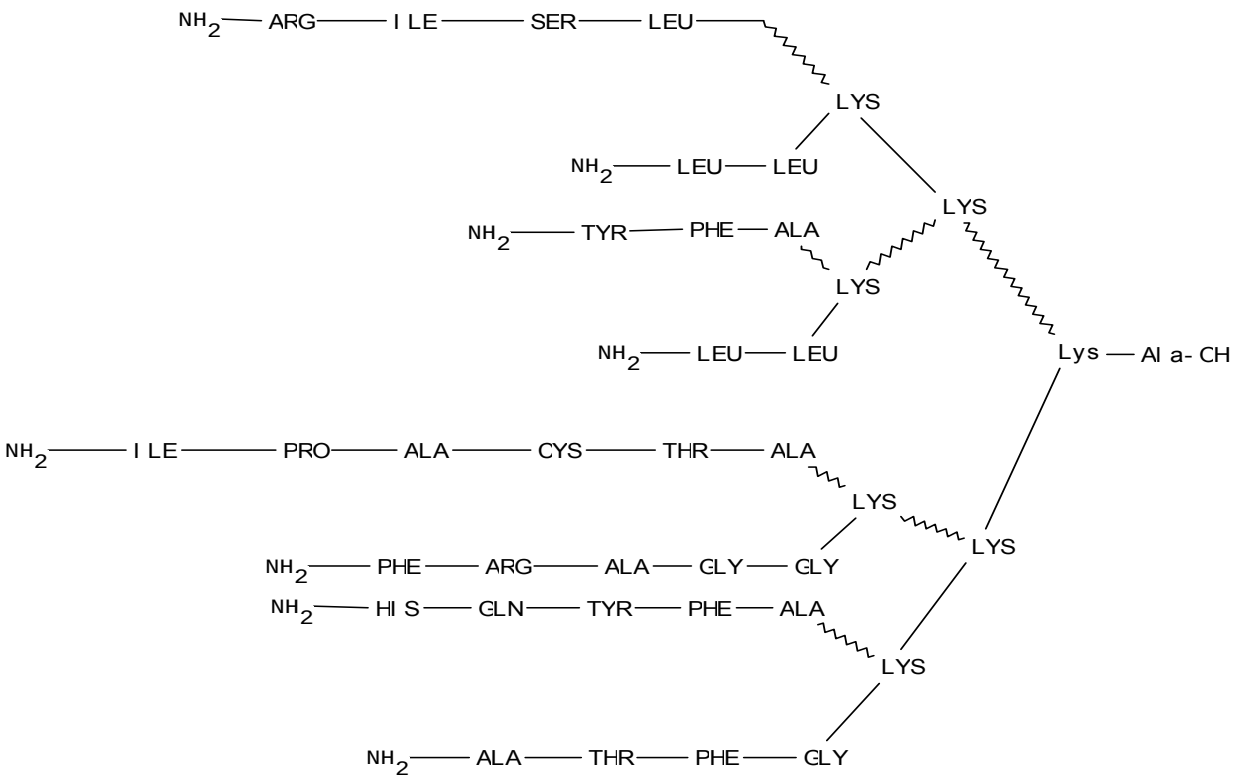
XXXXXXXXDXXXXXXXXXXFXXXXXXXXXXXXXXXXXXXXXXXXXXXXAXXXXXXX
XXXXXXXXXXXXGXXXXX (SEQ ID NO: 25)

Since "X" can be any amino acid, annotation of the "X" residues is not required under paragraph 26.

Relevant ST.26 paragraph(s): Paragraphs **6(b)**, 7 and 26

Example 6(b)-2: Branched amino acid sequence

The application describes a branched sequence where the Lysine residues are used as a scaffolding core to form eight branches to which multiple linear peptide chains are attached. Lysine is a dibasic amino acid, providing it with two sites for peptide-bonding. The peptide is illustrated as follows:



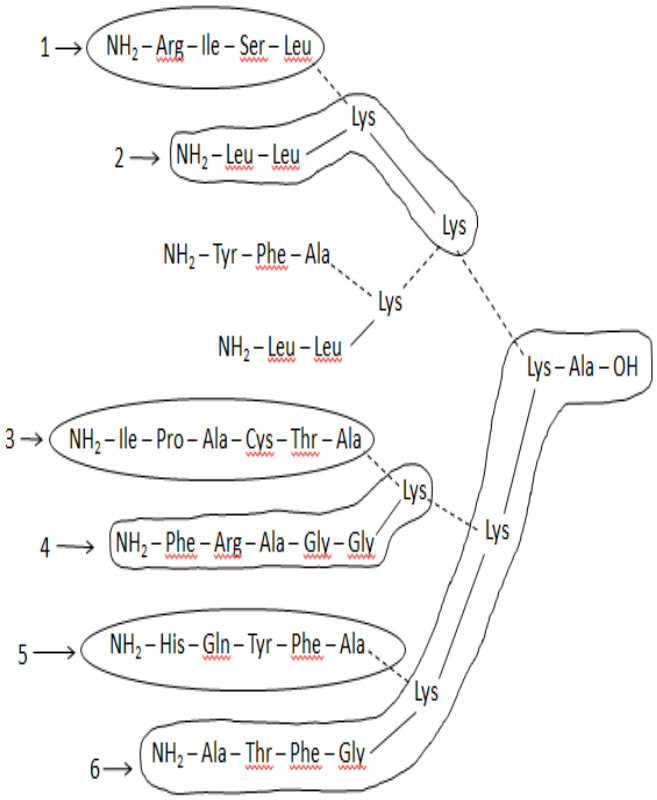
In the above branched peptide, the bonds depicted by —— represent an amide linkage between the terminal amine of the Lysine and the carboxyl end of the bonded amino acid. The bonds depicted by ∿∿∿ represent an amide linkage between the side chain amine of the Lysine and the carboxyl end the bonded amino acid.

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

YES

The example discloses a branched sequence where the lysine residues are used as a scaffolding. Paragraph 6(b) requires that the unbranched or linear portion of the sequence, containing four or more specifically defined amino acids, be included in a

sequence listing. In the above example, the linear portions of the branched peptide that have four or more amino acids are encircled:



ST.26 paragraph 6(b) requires inclusion of peptides 1-6 above in a sequence listing.

Peptides which are not required, and in fact are prohibited, from inclusion in the sequence listing are:

YFA
LLK

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

Peptides 1-6 must be represented with separate sequence identifiers:

RISL (SEQ ID NO: 26)

LLKK (SEQ ID NO: 27)

IPACTA (SEQ ID NO: 28)

FRAGGK (SEQ ID NO: 29)

HQYFA (SEQ ID NO: 30)

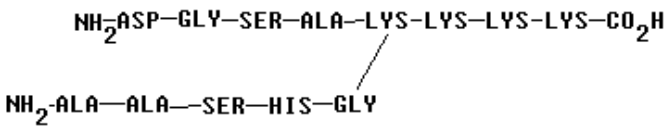
ATFGKKKA (SEQ ID NO: 31)

The cross linkage is preferably noted using the feature key “SITE” and the mandatory qualifier “NOTE” with the value e.g., “This sequence is one part of a branched peptide”.

Relevant ST.26 paragraph(s): Paragraphs **6(b)**, 25, 29, and 30

Example 6(b)-3: Branched amino acid sequence

Peptide of the following sequence:

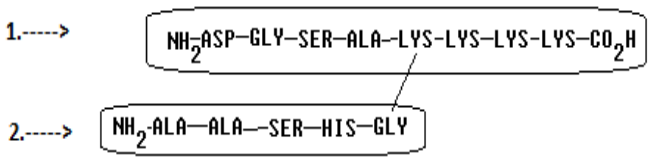


The linkage between the terminal Glycine residue in the lower sequence and the Lysine in the upper sequence is through a peptide bond between the carboxy terminus of the Glycine and the amino terminal side chain of the Lysine.

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

YES

The unbranched or linear portion of a sequence, containing four or more specifically defined amino acids, must be included in a sequence listing. In the above example, the linear portions of the branched peptide that have more than four amino acids are:



ST.26 paragraph 6(b) requires inclusion of peptides 1 and 2 in a sequence listing.

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

Peptides 1 and 2 must be represented with separate sequence identifiers:

DGSAKKK (SEQ ID NO: 32)

AASHG (SEQ ID NO: 33)

Preferably the sequence DSAKKKK should include an annotation to indicate that the 5th lysine is a modified amino acid using the feature key “SITE” together with the qualifier “NOTE” describing that lysine links the peptide AASHG. Preferably the

sequence AASHG should include an annotation to indicate that the 5th glycine is linked to DGSAKKK using the feature key “SITE” together with the qualifier “NOTE”.

Relevant ST.26 paragraph(s): Paragraphs **6(b)**, 25, 29, and 30

Paragraph 10(a) – Double-stranded nucleotide sequence – fully complementary

Example 10(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)?

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: Paragraphs 6(a), **10(a)**, and 13

Paragraph 10(b) – Double-stranded nucleotide sequence - not fully complementary

Example 10(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 – the ABC1 promoter region (top strand)

The top strand has more than ten enumerated and "specifically defined" nucleotides and is required to be included in a sequence listing.

YES – the PNA probe (bottom strand)

The bottom strand must also be included in the sequence listing, with its own sequence identification number, because the two strands are not fully complementary to each other. The individual residues that comprise a PNA or "peptide nucleic acid" are considered nucleotides according to ST.26 paragraph 3(d). 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 10, it must be included in a sequence listing "in the direction from left to right that mimics the 5'–end to 3'-end direction." Therefore, it must be included in a sequence listing as:

cgcctnagtcaatggg (SEQ ID NO: 37)

The “organism” qualifier of the feature key “source” must have the value “synthetic construct” and the mandatory qualifier “mol_type” with the value “other DNA”. The bottom strand must be described in a feature table using the feature key “modified_base” and the mandatory qualifier “mod_base” with the abbreviation “OTHER”. A “note” qualifier must be included with the complete unabbreviated name of the modified nucleotides, such as “N-(2-aminoethyl) glycine nucleosides”.

The “n” residue must be further described in a feature table using the feature key “modified_base” and the mandatory qualifier “mod_base” with the abbreviation “OTHER”. A “note” qualifier must be included with the complete unabbreviated name of the modified nucleotide: “N-(2-aminoethyl) glycine 5-nitroindole or N-(2-aminoethyl) glycine 3-nitroindole”.

Relevant ST.26 paragraphs: Paragraphs 3(d), 6(a), **10(b)**, 17, and 17bis

Example 10(b)-2: Double-stranded nucleotide sequence – no base-pairing segment

A patent application describes the following double-stranded DNA sequence:

```
3′-CCGGTTAGCTTATACGCTAGGGCTA-5′
   |||||||       ||||||||||||
5′-GGCCAATATGGCTTGCGATCCCGAT-3′
```

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

YES

Each strand of the enumerated, double-stranded nucleotide sequence has more than 10 specifically defined nucleotides. Both strands must be included in the sequence listing, each with its own sequence identification number, because the two strands are not fully complementary to each other.

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

The sequence of each strand must be represented in the 5’ to 3’ direction and assigned its own sequence identification number:

atcgggatcgcatattcgattggcc (top strand) (SEQ ID NO: 38)

and

ggccaatatggcttgcgatcccgat (bottom strand) (SEQ ID NO: 39)

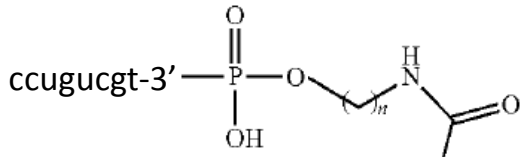
Relevant ST.26 paragraphs: Paragraphs 6(a), **10(b)**, and 13

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

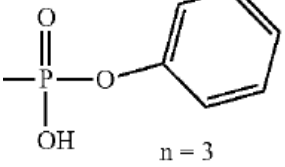
Example 14-1: The symbol “t” represents uracil in RNA

A patent application describes the following compound:

segment A: ccugucgt-3’



segment B: uaguuguagaggccugucct-5’



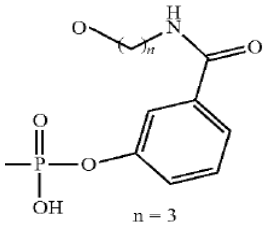
Wherein segment A and segment B are RNA sequences.

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(d); therefore, each segment must be considered a separate sequence. Segment B contains more than 10 specifically defined nucleotides and ST.26 paragraph 6(a) requires inclusion in a sequence listing. Segment A contains only 8 specifically defined nucleotides and therefore is not required to be included in a sequence listing.

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

Segment A contains fewer than 10 specifically defined nucleotides, and therefore it cannot 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” cannot 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., “ccugucgt (Segment A) is attached at its 3'-end to a linker which is attached to the 5’ oxygen of the thymidine. The linker is (4-(3-hydroxybenzamido)butyl) phosphinic acid.”

Relevant ST.26 paragraphs: Paragraphs 3(d), 6(a), 7, 13, **14**, 18, and 54

Paragraph 26 – The most restrictive ambiguity symbol should be used

Example 26-1: Shorthand formula for a nucleotide sequence

$(GGGz)_2$

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.

Therefore, "z" is equivalent to the conventional symbol "X", and the peptide in the example has eight enumerated amino acids, six of which are specifically defined glycine residues. ST.26 paragraph 6(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 6(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)

Further, the example does not disclose that "z" is the same amino acid in both positions in the expanded sequence. However, if "z" is disclosed as the same amino acid in both positions, then a feature key "VARIANT" and a qualifier "NOTE" should be provided stating that "X" in position 4 and 8 can be any amino acid, as long as they are the same in both positions.

Relevant ST.26 paragraph(s): Paragraphs 3(b*bis*), 6(b) and **26**

Example 26-2: Shorthand formula - less than four specifically defined amino acids

A peptide of the formula $(Gly\text{-}Gly\text{-}Gly\text{-}z)_n$

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 6(b) does not require inclusion, despite the fact that "n" is also defined as specific numerical values in some embodiments.

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

YES

The example provides a specific numerical value for variable "n," i.e., a lower limit of 2, an upper limit of 100, and an exact value 3. Any sequence containing at least four specifically defined amino acids may be included in the sequence listing.

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

A sequence containing 100 copies of GGGX is preferred (SEQ ID NO: 42). A further annotation should indicate that up to 98 copies of GGGX could be deleted. Inclusion

of further specific embodiments that are a key part of the invention is strongly encouraged.

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

Relevant ST.26 paragraph(s): Paragraphs 3(b*bis*), 6(b), 25, and **26**

Example 26-3: Shorthand formula - four or more specifically defined amino acids

A peptide of the formula $(Gly\text{-}Gly\text{-}Gly\text{-}z)_n$

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: 43) (*see* Introduction to this document). In the latter case, a further annotation could indicate that up to 98 copies of GGGX could be deleted. Inclusion of two additional sequences containing 2 and 3 copies of GGGX, respectively (SEQ ID NO: 44-45), is strongly encouraged.

Further, the example does not disclose that the "z" variable is the same in each of the two occurrences in the expanded sequence. However, if "z" is disclosed as the same amino acid in all locations, then a feature Key VARIANT and a Qualifier NOTE should indicate that "X" in all positions can be any amino acid, as long as they are the same in all locations.

CAUTION: The preferred representation of the sequence indicated above is directed to the provision of a sequence listing on the filing date of a patent application. The same representation may not be applicable to a sequence listing provided subsequent to the filing date of a patent application, since consideration

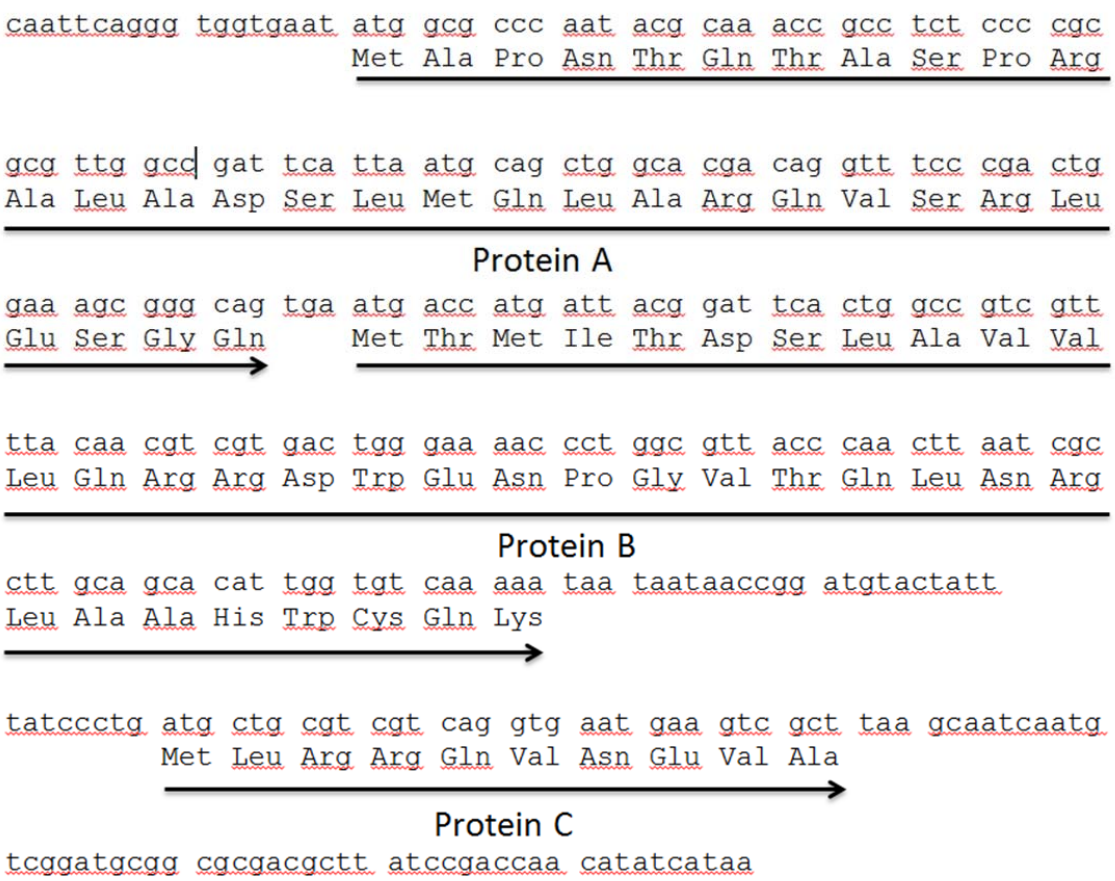
must be given to whether the information provided could be considered by an IPO to add subject matter to the original disclosure.

Relevant ST.26 paragraph(s): Paragraphs 3(b*bis*), 6(b), 25, and **26**

Paragraph 27 – Amino acid sequences separated by internal terminator symbols

Example 27-1: 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 acids sequences.

The enumerated nucleotide sequence contains more than 10 specifically defined nucleotides and must be included in a sequence listing as a single sequence.

Regarding the encoded amino acid sequences, paragraph 27 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 6(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:

caattcagggtggtgaatatggcgcccaatacgcaaaccgcctctccccgcgcgttggccgattcattaatgga
aagcgggcagtgaatgaccatgattacggattcactggccgtcgttttacaacgtcgtgactgggaaaaccctg
gcgttacccaacttaatcgccttgcagcacattggtgtcaaaaataataataaccggatgtactatttatccctgatg
ctgcgtcgtcaggtgaatgaagtcgcttaagcaatcaatgtcggatgcggcgcgacgcttatccgaccaacatat
cataa. (SEQ ID NO: 46)

The nucleotide sequence should further be described using a “CDS” feature key for each of the three proteins and the element INSDFeature_location should identify the location of each coding sequence, including the stop codon. In addition, for each “CDS” feature key, the “translation” qualifier should be included with the amino acid sequence of the protein as the qualifier value. The application does not disclose the genetic code table that applies to the translation (see Annex 1, Section 9, Table 5). If the Standard Code table applies, then the qualifier “transl_table” is not necessary; however, if a different genetic code table applies, then the appropriate qualifier value from Table 5 must be indicated for the qualifier “transl_table”. Finally, the qualifier “protein _id” must be included with the qualifier value indicating the sequence identification number of each of the translated amino acid sequences.

The amino acid sequences must be included as separate sequences, each assigned its own sequence identification number:

MAPNTQTASPRALADSLMQLARQVSRLESGQ (SEQ ID NO: 47)

MTMITDSLAVVLQRRDWENPGVTQLNRLAAHWCQK (SEQ ID NO: 48)

MLRRQVNEVA (SEQ ID NO: 49)

Relevant ST.26 paragraphs: Paragraphs 6, 25, **27**, 57, 88-90

Paragraph 28 – Representation of an "other" amino acid

Example 28-1: Most restrictive ambiguity symbol for an "other" amino acid

A patent application describes the following sequence:

Ala-Hse-X_1-X_2-X_3-X_4-Tyr-Leu-Gly-Ser

Wherein, X_1= Ala or Gly,

X_2= Ala or Gly,

X_3= Ala or Gly,

X_4= 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 6(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 28 requires any "other" amino acid must be represented by the symbol "X". In the example, the sequence contains the amino acid Hse in position 2 which is not found in Annex I, Section 3, Table 3. Accordingly, Hse is an "other" amino acid and must be represented by the symbol "X".

X_1-X_4 are variant positions, each of which can be A or G. The most restrictive ambiguity symbol for alternatives A or G is "X". Therefore, the sequence may be represented as:

AXXXXXYLGS (SEQ ID NO: 50)

Inclusion of any specific sequences essential to the disclosure or claims of the invention is highly recommended, as discussed in the introduction to this document.

Since amino acid Hse is not found in Annex I, Section 4, Table 4, a feature key “SITE” and a qualifier “NOTE” must be provided with the complete, unabbreviated name of Homoserine.

According to paragraph 26, because X_1-X_4 represent an alternative of only 2 amino acids, then further description is required. Paragraph 93 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 can be jointly described using the syntax “3..6” as the location descriptor in the element INSDFeature_location.

Relevant ST.26 paragraphs: Paragraphs 3(a), 6(b), 24-26, **28**, 34, 67, 71, 72, and 93-94

Paragraph 29 – Annotation of a modified amino acid

Example 29-1 – 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 (6)(b).

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

According to ST.26 paragraph 28, a modified amino acid should be represented in the sequence as the corresponding unmodified amino acid whenever possible.

Therefore the sequence must be included in a sequence listing as:

LEYCLKRWNETISHCAW (SEQ ID NO: 51)

A further description of the modified amino acid is required. The feature key "CARBOHYD" together with the (mandatory) qualifier "NOTE" should be used to indicate the occurrence of the attachment of a sugar chain (asialyloligosaccharide) to asparagine in position 9. The qualifier "NOTE" describes the type of linkage, e.g. N-linked. The location descriptor in the feature location element is the residue number of the modified asparagine.

In addition, there is a disulfide bond between the two Cys residues. Therefore the feature key "DISULFID" is used to describe an intrachain crosslink. The location descriptors in the feature location element are the residue numbers of the linked Cys residues in conjunction with the "join" location operator, "join(4,15)". The qualifier NOTE is not mandatory.

Relevant ST.26 paragraph(s): Paragraphs 3(a), 6(b), 25, 28, **29**, and Annex I, section 7, feature key 7.4

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

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

LL-100-KYMR

Where the “-100-“between amino acids Leucine and Lysine reflects a 100 amino acid region in the sequence.

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

YES

ST.26 paragraph 36 requires inclusion of a sequence that contains at least four specifically defined amino acids separated by one or more regions of a defined number of “X” residues.

The disclosed sequence uses a nonconventional symbol, i.e. “-100-.” The definition of “-100-” must be determined from the explanation of the sequence in the disclosure, which defines this symbol as 100 amino acids between leucine and lysine (*see* Introduction to this document). Therefore, “-100-” is a defined region of “X” residues. Since six of the 106 amino acids in the sequence are specifically defined, ST.26 paragraph 6(b) requires that the sequence must be included in a sequence listing.

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

The nonconventional symbol “-100-” is represented as 100 “X” residues (since any symbol used to represent an amino acid is equivalent to only one residue). Therefore, a single sequence of 106 amino acids in length, containing 100 “X” residues between LL and KYMR, must be included in a sequence listing (SEQ ID NO: 52).

Relevant ST.26 paragraph(s): Paragraphs 6(b), 25, 26, and **36**

Example 36-2: Sequence with multiple regions of a known number or range of "X" residues represented as a single sequence

Lys-z_2-Lys-z_m-Lys-z_3-Lys-z_n-Lys-z_2-Lys

Where z is any amino acid, m=20, n=19-20, z_2 means that the pairs of Lysines are separated by any two amino acids, and z_3 means the pairs of Lysines are separated by any three amino acids.

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

YES

The disclosed sequence uses a nonconventional symbol, i.e. "z." Therefore, the disclosure must be consulted to determine the definition; "z" is defined as any amino acid (*see* Introduction to this document). The conventional symbol used to represent any amino acid is "X". Considering the presence of "X" variables, the peptide contains six lysine residues that are enumerated and specifically defined, which is required to be included in a sequence listing.

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

The sequence uses a nonconventional symbol "z", the definition of which must be determined from the disclosure. Since "z" is defined as any amino acid, the conventional symbol is "X."

The preferred and most encompassing means of representation is (*see* Introduction to this document):

KXXKXXXXXXXXXXXXXXXXXXXXKXXXKXXXXXXXXXXXXXXXXXXXXKXXK (SEQ ID NO: 53)

wherein z_n is equal to 20 "X's", with a further description that the "X" variable corresponding to position 30 can be deleted.

Alternatively, or in addition to the above, the sequence can be represented as:

KXXKXXXXXXXXXXXXXXXXXXXXKXXXKXXXXXXXXXXXXXXXXXXXXKXXK (SEQ ID NO: 54)

wherein z_n is equal to 19 "X's", with a further description that an "X" variable between positions 29 and 30 can be inserted.

Relevant ST.26 paragraph(s): Paragraphs 25, 26, and **36**

Example 36-3: Sequence with multiple regions of a known number or range of "X" residues represented as a single sequence

K-z_2-K-z_m-K-z_3-K-z_n-K-z_2-K

Where z is any amino acid , where m=15-25, preferably 20-22, n=15-25, preferably 19-20, z_2 means that the pairs of Lysines are separated by any two amino acids, and z_3 means the pairs of Lysines are separated by any three amino acids.

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

YES

The sequence in the example uses a nonconventional symbol, i.e. "z." Therefore, the surrounding disclosure is consulted to determine the definition of "z" (*see* Introduction to this document). The disclosure defines this symbol as any amino acid. The conventional symbol used to represent this amino acid is "X." After considering the presence of "X" variables, the peptide contains 6 lysine residues that are enumerated and specifically defined, which is required in a sequence listing.

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

The sequence uses a nonconventional symbol "z", the definition of which must be determined from the disclosure. Since "z" is defined as any amino acid, the conventional symbol is "X". The preferred and most encompassing means of representation is:

KXXKXXXXXXXXXXXXXXXXXXXXXXXKXXXKXXXXXXXXXXXXXXXXXXXXXXXXXXXKXXK (SEQ ID NO: 55)

(where m=25 and n=25), with a further description that up to 10 "X" residues in each of the "z_m" or "z_n" regions may be deleted.

Inclusion of any specific sequences essential to the disclosure or claims of the invention is highly recommended, as discussed in the introduction to this document.

Alternatively, the sequence can be represented as:

KXXKXXXXXXXXXXXXXXKXXXKXXXXXXXXXXXXXXXKXXK (SEQ ID NO: 56)

(where m=15 and n=15), with a further description that up to 10 "X" residues in each of the "z_m" or "z_n" regions may be inserted.

As further alternatives, any or all possible variations may be included.

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

Relevant ST.26 paragraph(s): Paragraphs 26 and **36**

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

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

Gly-Gly----Gly-Gly-Xaa-Xaa

where the symbol ---- is an undefined gap within the sequence, where Xaa is any amino acid, and the Glycine and Xaa residues are connected to one another through peptide bonds.

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

NO

ST.26 paragraph 37 prohibits the inclusion of any sequence that contains an undefined gap; therefore, inclusion of the entire sequence is not required.

ST.26 paragraph 37 does require inclusion of any portion of a sequence adjacent to an undefined gap that contains four or more specifically defined amino acids. In the example above, inclusion of either portion adjacent to the undefined gap is not required, since each portion contains only two specifically defined amino acids.

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

NO – not the entire sequence

NO – not any portion of the sequence

ST.26 paragraph 37 does not permit inclusion of the entire sequence.

ST.26 paragraph 7 does not permit inclusion of either portion adjacent to the undefined gap, since each portion contains only two specifically defined amino acids.

Relevant ST.26 paragraphs: Paragraphs 6(b), 7, 25, and **37**

Example 37-2: Sequence with regions of an unknown number of "X" residues must not be represented as a single sequence

Gly-Gly----Gly-Gly-Ala-Gly-Xaa-Xaa

wherein the symbol ---- is an undefined gap within the sequence, where Xaa is any amino acid, and the Glycine and Xaa residues are connected to one another through peptide bonds.

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

NO – not the entire sequence

YES – a portion of the sequence

ST.26 paragraph 37 prohibits the inclusion of any sequence that contains an undefined gap, but requires inclusion of any portion of a sequence adjacent to an undefined gap that contains four or more specifically defined amino acids.

In the example above, ST.26 does not require (and prohibits) inclusion of both the entire sequence, which contains an undefined gap, and the Gly-Gly portion adjacent to the undefined gap, which contains only two specifically defined amino acids. However, ST.26 requires inclusion of the Gly-Gly-Ala-Gly- Xaa-Xaa portion adjacent to the undefined gap, since it contains at least four specifically defined amino acids.

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

NO – not the entire sequence and not the Gly-Gly portion

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

The portion of the sequence adjacent to the undefined gap that contains four specifically defined amino acids should be represented as:

GGAGXX (SEQ ID NO: 57)

Preferably, the sequence should be annotated to indicate that the represented sequence is part of a larger sequence that contains an undefined gap by using the feature key "SITE", the feature location "1" and the qualifier "NOTE" with the value, e.g., "This residue is linked N-terminally to a peptide having an N-terminal Gly-Gly and a gap of undefined length.".

Relevant ST.26 paragraph(s): Paragraphs 6(b), 7, 25, and **37**

Paragraph 88 – "CDS" Feature key

Example 88-1: 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 nucleotide sequence.** The enumerated nucleotide sequence has more than ten specifically defined nucleotides.

YES – **the peptide sequence.** The enumerated peptide sequence has more than four specifically defined amino acids.

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

The nucleotide sequence must be presented as:

atgaccggaaataaacctgaaaccgatgtttacgaaattttatga (SEQ ID NO: 58)

The nucleotide sequence should further be described using the "CDS" feature key and the element INSDFeature_location should identify the entire sequence, including the stop codon (i.e., position 1 through 45). In addition, the "translation" qualifier should be included with the qualifier value "MTGNKPETDVYEIL". The application does not disclose the genetic code table that applies to the translation (see Annex 1, Section 9, Table 5). If the Standard Code table applies, then the qualifier "transl_table" is not necessary; however, if a different genetic code table applies, then the appropriate qualifier value from Table 5 must be indicated for the qualifier "transl_table". Finally, the qualifier "protein _id" must be included with the qualifier value indicating the sequence identification number of the translated peptide.

The peptide sequence must be separately presented with its own sequence identification number using single letter codes as follows:

MTGNKPETDVYEIL (SEQ ID NO: 59)

The STOP following the enumerated peptide sequence must not be included in the peptide sequence in the sequence listing.

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

Relevant ST.26 paragraphs: Paragraphs 6(a), 6(b), 25, 27, **88**, 89, and 90

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

Example 91-1: Representation of enumerated variants

The description includes the following sequence alignment.

```
D. melanogasterACATTGAATCTCATACCACTTT
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 species 2 and 3 are considered enumerated and specifically defined nucleotides, as they are simply a short-hand way of indicating that a given position is the same nucleotide as in species 1. In addition, sequence alignments frequently display the symbol “-“ to indicate the absence of a residue in order to maximize the alignment.

Accordingly, the nucleotide sequences of species 1 and 3 contain twenty-two enumerated and specifically defined nucleotides, whereas the nucleotide sequences of species 2 contains nineteen. Thus, each sequence is required by ST.26 paragraph 6(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: 60)

Drosophila virilis sequence must be included in a sequence listing as: acatggatcccacgacttt (SEQ ID NO: 61)

Drosophila simulans sequence must be included in a sequence listing as: gtatggcgtcgtatsgtagttt (SEQ ID NO: 62)

Relevant ST.26 paragraphs: Paragraphs 6(a), 13, and **91**

Paragraph 91*bis* – Variant sequence disclosed as a single sequence with enumerated alternative residues

Example 91*bis*-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 6(b) requires inclusion of the sequence in a sequence listing.

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

Table 3 of Annex I, Section 3 defines the ambiguity symbol "J" as isoleucine or leucine. Therefore, the preferred representation of the sequence is:

GGGJATX (SEQ ID NO: 63)

which requires a further description in a feature table using the feature key "VARIANT" and the qualifier "NOTE" to indicate that the "X" is Serine or Threonine.

Alternatively, the sequence could be represented as:

GGGLATS (SEQ ID NO: 64)

which requires a further description in a feature table using the feature key "VARIANT" and the qualifier "NOTE" to indicate that L can be replaced by I, and S can be replaced by T.

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

Relevant ST.26 paragraph(s): Paragraphs 6(b), 7, 25, 26, **91*bis***, and 94

Paragraph 92(a) – A variant sequence disclosed only by reference to a primary sequence with multiple independent variations

Example 92(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 6(b).

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

In this example, the enumerated sequence of "Peptide fragment 1" is disclosed three times, as three different embodiments, each with an alternative description of Xaa. In this example, "X" is the most restrictive ambiguity symbol for the Xaa position.

ST.26 requires inclusion of the disclosed enumerated sequence only once. In the most encompassing of the three embodiments, Xaa is any amino acid (*see* Introduction to this document). Therefore, the sequence that must be included in the sequence listing is:

GLPXRIC (SEQ ID NO: 65)

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

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

GLPVRIC (SEQ ID NO: 66)
GLPTRIC (SEQ ID NO: 67)
GLPDRIC (SEQ ID NO: 68)

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

Relevant ST.26 paragraph(s): Paragraphs 6(b), 25, 26, and **92(a)**

Paragraph 92(b) – A variant sequence disclosed only by reference to a primary sequence with multiple interdependent variations

Example 92(b)-1: Representation of individual variant sequences with multiple interdependent variations

A patent application describes the following consensus sequence:

cgaatg**n$_1$**cccactacgaatg**n$_2$**cacgaatg**n$_3$**cccaca

wherein n_1, n_2, and n_3 can be a, t, g, or c.

Several variant sequences are disclosed as follows:

if n_1 is a, then n2 and n3 are t, g, or c;

if n_1 is t, then n2 and n3 are a, g, or c;

if n_1 is g, then n2 and n3 are t, a, or c;

if n_1 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 6(a) to be included in a sequence listing.

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

The enumerated sequence contains more than ten specifically defined nucleotides and three "n" residues. ST.26 requires inclusion of the disclosed enumerated sequence and where an ambiguity symbol is appropriate, the most restrictive symbol should be used. In this example, n_1, n_2, and n_3 can be a, t, g, or c, so "n" is the most restrictive ambiguity symbol. Therefore, the sequence that must be included in the sequence listing is:

cgaatgncccactacgaatgncacgaatgncccaca (SEQ ID NO: 69)

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 92(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: 70)

cgaatgtcccactacgaatgvcacgaatgvcccaca (SEQ ID NO: 71)

cgaatggcccactacgaatghcacgaatghcccaca (SEQ ID NO: 72)

cgaatgccccactacgaatgdcacgaatgdcccaca (SEQ ID NO: 73)

(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 cannot be represented by "n" in the sequence.

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

Relevant ST.26 paragraphs: Paragraphs 6(a), 15, and **92(b)**