
UNITED STATES
PATENT AND TRADEMARK OFFICE
uspto

The meat inside the WIPO Standard ST.26 nutshell

The details of sequence listings in patent applications

Susan Wolski

February 2019

UNITED STATES PATENT AND TRADEMARK OFFICE

uspto

WIPO Standard ST.26 and its annexes

- The Main Body – Requirements for inclusion/representation
- Annex I – Controlled vocabulary
- Annex II – Document Type Definition for Sequence Listing (DTD)
- Annex III – Sequence Listing Specimen (XML file)
- Annex IV – Character Subset from the Basic Latin Code Table for Use in an XML Instance of a Sequence Listing
- Annex V – Data Exchange Requirements (Patent Offices Only)
- Annex VI – the Guidance Document
  - Appendix – Guidance Document Sequences in XML
- Annex VII – Recommendation for the Transformation of a Sequence Listing from ST.25 to ST.26

ST.26 main body - expanded scope

Terminology	In ST.26 - Expanded to Include
Amino acid	• D-amino acids • Amino acids containing modified or synthetic side chains
Enumeration of its residues	Residues represented by name, abbreviation, symbol, or structure
Nucleotide	Nucleotide analogs, e.g., PNA, GNA, LNA, TNA
Sequences - required in a sequence listing	A linear region of a branched sequence containing - 10 or more specifically defined nucleotides, or - 4 or more specifically defined amino acids

ST.26 main body: sequence listing (SL) structure

- Presented in XML 1.0 using ST.26 Annex II DTD
- Contained in one Unicode UTF-8 file
  - General information part
    - Information on the patent application to which the SL is directed
    - Any Unicode characters, except reserved characters
  - Sequence data part
    - One or more sequence data elements, each containing information about one sequence
    - Printable characters from the Unicode Basic Latin code table
      - including the space character
      - excluding reserved characters

ST.26 main body: sequence data element for one sequence

Element	Description
INSDC_length	Length of the sequence
INSDC_moltype	Molecule type
INSDC_division	Indication that sequence is related to a patent application
INSDC_feature-table	List of annotations of the sequence
INSDC_sequence	Sequence

ST.26 sequence example

```
<SequenceData sequenceIDNumber="4">
    <INSDSeq>
        <INSDSeq_length>19</INSDSeq_length>
        <INSDSeq_moltype>AA</INSDSeq_moltype>
        <INSDSeq_division>PAT</INSDSeq_division>
        <INSDSeq_feature-table>
            <INSDFeature>
                <INSDFeature_key>SOURCE</INSDFeature_key>
                <INSDFeature_location>1..19</INSDFeature_location>
                <INSDFeature_quals>
                    <INSDQualifier>
                        <INSDQualifier_name>ORGANISM</INSDQualifier_name>
                        <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
                    </INSDQualifier>
                    <INSDQualifier>
                        <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
                        <INSDQualifier_value>protein</INSDQualifier_value>
                    </INSDQualifier>
                </INSDFeature_quals>
            </INSDFeature>
        </INSDSeq_feature-table>
        <INSDSeq_sequence>MLAPDCPFDPTRIYSSSLC</INSDSeq_sequence>
    </INSDSeq>
</SequenceData>
```

ST.26 main body – feature table

- A feature table consists of one or more feature elements
- A feature element describes one feature, consisting of
  - A feature key – word or abbreviation describing the feature
  - A feature location – corresponding region of the sequence
  - Feature qualifiers – auxiliary information about a feature

ST.26 main body – feature keys

- ST.26 Annex I – feature keys
  - Section 5 – Nucleotide sequences (49)
  - Section 7 – Amino acid sequences (39)
- Mandatory feature keys
  - “source” for nucleotide sequences
  - “SOURCE” for amino acid sequences

ST.26 main body – feature location: at least one location descriptor

Location Descriptor type	Syntax
Single residue number	x
Residue numbers delimiting a sequence span	x..y
Residues before the first or beyond the last specified residue number	<x >x <x..y x..>y
A site between two residue numbers	x^y

ST.26 main body – feature location: location operator

Location operators for nucleotides and amino acids

Location Syntax	Location Description
join(location, location,...location)	Indicated locations are joined (end-to-end) to form one sequence
order(location, location,...location)	Elements found in specified order - nothing implied as to whether reasonable to join

Location operator for nucleotides only

Location Syntax	Location Description
complement(location)	Indicates feature located on complementary strand to that specified by the location descriptor

ST.26 main body – feature qualifiers

- ST.26 Annex I – qualifiers and value formats
  - Section 6 – Nucleotide sequences (80)
  - Section 8 – Amino acid sequences (3)
- Mandatory feature qualifiers
  - Mandatory feature key “source”/“SOURCE” requires two mandatory qualifiers
    - “organism”/”ORGANISM” must disclose source of sequence
    - “mol_type”/”MOL_TYPE” must disclose type of molecule
  - Some optional feature keys require mandatory qualifiers

ST.26 main body – qualifier value format

- Free text is a type of value format for certain qualifiers
- Descriptive text phrase *preferably* in English
  - UTF-8 Unicode Basic Latin characters only
  - May require translation for National/Regional procedures
- Must not exceed 1000 characters

```
<INSDFeature>
  <INSDFeature_key>REGION</INSDFeature_key>
  <INSDFeature_location>&gt;5</INSDFeature_location>
  <INSDFeature_quals>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>The entire sequence of amino acids 1-5 can be repeated one or more
times</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>
```

ST.26 main body – coding sequences

- "CDS" feature key identifies coding sequences
  - Mandatory location must include stop codon
  - "transl_table" and "translation" qualifiers may accompany "CDS"
  - Standard Code Table assumed if "transl_table" qualifier not used
  - Amino acid sequences disclosed in a "translation" qualifier and encompassed by paragraph 7 must be included in the SL and assigned its own SEQ ID NO

ST.26 main body – variant sequences (1)

- Primary sequence and any variants disclosed by enumeration of their residues
  - Must each be included with their own SEQ ID NO
  - *See* ST.26 paragraph 91
- A variant disclosed as a single sequence with enumerated alternatives at 1 or more positions
  - Should be included as a single sequence with alternatives represented by most restrictive ambiguity symbol
  - *See* ST.26 paragraph 92

ST.26 main body – variant sequences (2)

- A variant disclosed only by reference to deletion(s), insertion(s), or substitution(s) in a primary sequence
  - May be represented by annotation of the primary sequence where occurrence of variations are independent
  - Should be represented as a separate sequence where occurrence of variations are interdependent
  - Must be represented as a separate sequence where it contains an inserted or substituted sequence in excess of 1000 residues
  - *See* ST.26 paragraph 93

ST.26 Annex VI – Guidance Document (1)

- Purpose
  - To ensure all applicants and IPOs understand and agree on the requirements for inclusion and representation of sequence disclosures
- Introduction provides further explanation of
  - Enumeration of its residues
  - Specifically defined
  - Most encompassing sequence
  - Proper usage of ambiguity symbols “n” and “X”

ST.26 Annex VI – Guidance Document (2)

- Each of the 48 examples considers three questions
  - Does ST.26, paragraph 7 require inclusion?
  - If not required, is inclusion permitted?
  - If inclusion is required or permitted, how should the sequence be represented?
- Each example is numbered according to the Main Body paragraph illustrated
- Sequences from each example are identified with a SEQ ID NO and included in the XML SL Annex

ST.26 paragraph 3(a) definition of “amino acid”

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

(a) “amino acid” means any amino acid that can be represented using any of the symbols set forth in Annex I (see Section 3, Table 3). Such amino acids include, inter alia, D-amino acids and amino acids containing modified or synthetic side chains. Amino acids will be construed as unmodified L-amino acids unless further described in the feature table as modified according to paragraph 30. For the purpose of this standard, a peptide nucleic acid (PNA) residue is not considered an amino acid, but is considered a nucleotide as set forth in paragraph 3(g)(i)(2).

ST.26 Annex VI – Example 3(a)-1

Example 3(a)-1: D amino acids

A patent application describes the following sequence:

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

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

YES

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

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

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

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

Paragraph 25 indicates that when amino acid sequences are circular in configuration and the ring consists solely of amino acid residues linked by peptide bonds, applicant must choose the amino acid in residue position number 1. Accordingly, the sequence may be represented as:

AEKXGMX (SEQ ID NO: 1)

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

Relevant ST.26 paragraphs: 3(a), 7(b), 25, 26, 29, 30, and 31

ST.26 Annex VI Appendix SEQ ID NO: 1 (1)

```
<SequenceData sequenceIDNumber="1">
  <INSDSeq>
    <INSDSeq_length>7</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..7</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
```

ST.26 Annex VI Appendix SEQ ID NO: 1 (2)

```
<INSDFeature>
  <INSDFeature_key>REGION</INSDFeature_key>
  <INSDFeature_location>1..7</INSDFeature_location>
  <INSDFeature_quals>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>circular peptide</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>1</INSDFeature_location>
  <INSDFeature_quals>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>D-Alanine</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>2</INSDFeature_location>
  <INSDFeature_quals>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>D-Glutamic acid</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>
```

ST.26 Annex VI Appendix SEQ ID NO: 1 (3)

```
<INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>4</INSDFeature_location>
  <INSDFeature_quals>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>Nle</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>6</INSDFeature_location>
  <INSDFeature_quals>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>D-Methionine</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>7</INSDFeature_location>
  <INSDFeature_quals>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>D-Norleucine</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>AEKXGMX</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
```

ST.26 paragraph 3(c) definition of “enumeration of its residues”

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

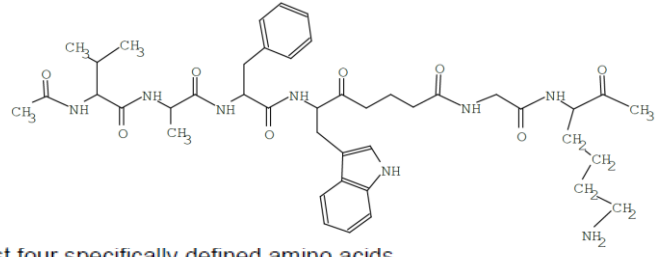
(i) the residue is represented by a name, abbreviation, symbol, or structure (e.g., HHHHHHQ or HisHisHisHisHisHisGln); or

(ii) multiple residues are represented by a shorthand formula (e.g., His_6Gln).

ST.26 Annex VI – Example 3(c)-1

Paragraph 3(c) – Definition of "enumeration of its residues"

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



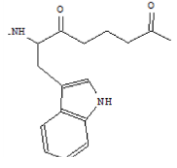
Question 1: Does ST.26 require inclusion of the sequence(s)?

YES

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

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

The sequence may be represented as:



VAFXGK (SEQ ID NO: 2)

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

Alternatively, the sequence may be represented as:

VAFW (SEQ ID NO: 3)

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

Relevant ST.26 paragraph(s): 3(c), 7(b), 29, 30, and 31

ST.26 Annex VI Appendix SEQ ID NO: 2 (1)

```
<SequenceData sequenceIDNumber="2">
  <INSDSeq>
    <INSDSeq_length>6</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..6</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>SITE</INSDFeature_key>
        <INSDFeature_location>1</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>the N-terminus is acetylated</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
```

ST.26 Annex VI Appendix SEQ ID NO: 2 (2)

```
        <INSDFeature>
          <INSDFeature_key>SITE</INSDFeature_key>
          <INSDFeature_location>4</INSDFeature_location>
          <INSDFeature_quals>
            <INSDQualifier>
              <INSDQualifier_name>NOTE</INSDQualifier_name>
              <INSDQualifier_value>6-amino-7-(1H-indol-3-yl)-5-oxoheptanoic acid</INSDQualifier_value>
            </INSDQualifier>
          </INSDFeature_quals>
        </INSDFeature>
        <INSDFeature>
          <INSDFeature_key>SITE</INSDFeature_key>
          <INSDFeature_location>6</INSDFeature_location>
          <INSDFeature_quals>
            <INSDQualifier>
              <INSDQualifier_name>NOTE</INSDQualifier_name>
              <INSDQualifier_value>the C-terminus is methylated</INSDQualifier_value>
            </INSDQualifier>
          </INSDFeature_quals>
        </INSDFeature>
      </INSDSeq_feature-table>
      <INSDSeq_sequence>VAFXGK</INSDSeq_sequence>
    </INSDSeq>
  </SequenceData>
```

ST.26 Annex VI Appendix SEQ ID NO: 3 (1)

```
<SequenceData sequenceIDNumber="3">
  <INSDSeq>
    <INSDSeq_length>4</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..4</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
```

ST.26 Annex VI Appendix SEQ ID NO: 3 (2)

```
<INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>1</INSDFeature_location>
  <INSDFeature_quals>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>the N-terminus is acetylated</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>4</INSDFeature_location>
  <INSDFeature_quals>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>C-terminus linked via a glutaraldehyde bridge to dipeptide
GK</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_quals>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>VAFW</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
```

ST.26 paragraph 3(g) definition of "nucleotide"

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

(i) a backbone moiety selected from:

(1) 2' deoxyribose 5' monophosphate (the backbone moiety of a deoxyribonucleotide) or ribose 5' monophosphate (the backbone moiety of a ribonucleotide); or

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

and

(ii) the backbone moiety is either:

(1) joined to a nucleobase, including a modified or synthetic purine or pyrimidine nucleobase; or

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

ST.26 Annex VI – Example 3(g)-3

Example 3(g)-3: Abasic site

A patent application describes the following sequence:

gagcattgac-AP-taaggct

Wherein AP is an abasic site

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

YES

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

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

The sequence must be included in a sequence listing as:

gagcattgacntaaggct (SEQ ID NO: 10)

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

Relevant ST.26 paragraphs: 3(g), 7(a), and 17

ST.26 Annex VI Appendix SEQ ID NO: 10

```
<SequenceData sequenceIDNumber="10">
  <INSDSeq>
    <INSDSeq_length>18</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..18</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>organism</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>mol_type</INSDQualifier_name>
            <INSDQualifier_value>genomic DNA</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>modified_base</INSDFeature_key>
        <INSDFeature_location>11</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>mod_base</INSDQualifier_name>
            <INSDQualifier_value>OTHER</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>abasic site</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>gagcattgacntaaggct</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
```

ST.26 paragraph 3(g) definition of "nucleotide"

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

(i) a backbone moiety selected from:

(1) 2' deoxyribose 5' monophosphate (the backbone moiety of a deoxyribonucleotide) or ribose 5' monophosphate (the backbone moiety of a ribonucleotide); or

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

and

(ii) the backbone moiety is either:

(1) joined to a nucleobase, including a modified or synthetic purine or pyrimidine nucleobase; or

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

ST.26 Annex VI – Example 3(g)-4

Example 3(g)-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(g)(i)(2). Accordingly, the sequence has more than ten enumerated and "specifically defined" nucleotides and is required to be included in a sequence listing.

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

GNA sequences do not have a 5'-end and a 3'-end, but rather, a 3'-end and a 2'-end. The 3'-end, which is routinely depicted as having a terminal phosphate group, corresponds to the 5'-end of DNA or RNA. (Note that other nucleic acid analogues may correspond differently to the 5'-end and 3'-end of DNA and RNA.) According to paragraph 11, it must be included in a sequence listing "in the direction from left to right that mimics the 5'-end to 3'-end direction." Therefore, it must be included in a sequence listing as:

tagttcattgactaaggctccccattgact (SEQ ID NO: 11)

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

Relevant ST.26 paragraphs: 3(d), **3(g)**, 7(a), 11, 16, 18, 65, and 66

ST.26 Annex VI Appendix SEQ ID NO: 11

```
<SequenceData sequenceIDNumber="11">
  <INSDSeq>
    <INSDSeq_length>30</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..30</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>organism</INSDQualifier_name>
            <INSDQualifier_value>synthetic construct</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>mol_type</INSDQualifier_name>
            <INSDQualifier_value>other DNA</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>GNA sequence</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>modified_base</INSDFeature_key>
        <INSDFeature_location>1..30</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>mod_base</INSDQualifier_name>
            <INSDQualifier_value>OTHER</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value> 2,3-dihydroxypropyl nucleosides (glycol nucleic acids)</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>tagttcattgactaaggctccccattgact</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
```

ST.26 paragraph 7 – sequences required in a SL

7. For the purpose of this Standard, a sequence for which inclusion in a sequence listing is required is one that is disclosed anywhere in an application by enumeration of its residues and can be represented as:

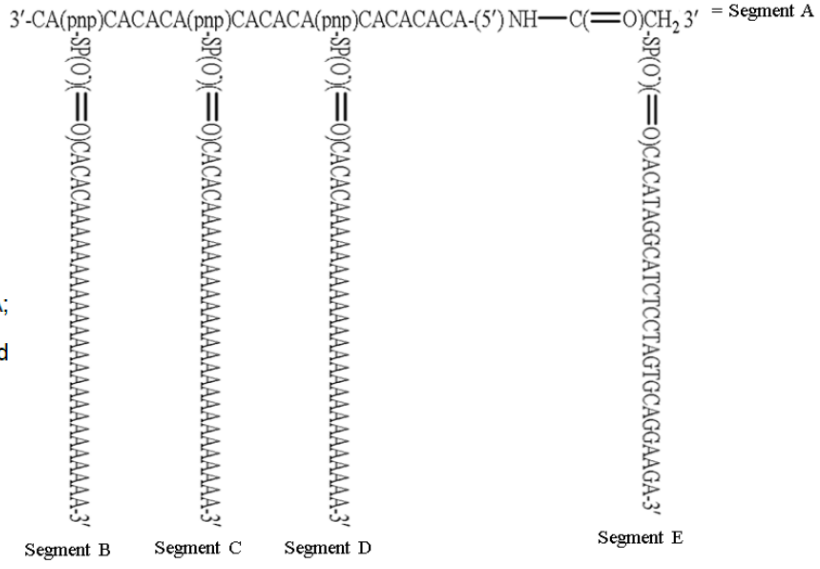
(a) an unbranched sequence or a linear region of a branched sequence containing ten or more specifically defined nucleotides, wherein adjacent nucleotides are joined by:

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

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

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

ST.26 Annex VI – Example 7(a)-1 (1)



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

Example 7(a)-1: Branched nucleotide sequence

The description discloses the following branched nucleotide sequence:

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

3'-CA(pnp)CACACA(pnp)CACACA(pnp)CACACACA-(5')NH—C(=O)CH_2 3' is segment A;

SP(O^-)(=O)CACACAAAAAAAAAAAAAAAAAAAAAAAA 3' is segments B, C, and D; and

SP(O^-)(=O)CACATAGGCATCTCCTAGTGCAGGAAGA 3' is segment E.

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

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

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

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

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

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

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

ST.26 Annex VI – Example 7(a)-1 (2)

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

NO

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

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

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

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

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

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): 7(a), 8, 11, 13, and 17

ST.26 Annex VI Appendix SEQ ID NO: 18

```
<SequenceData sequenceIDNumber="18">
  <INSDSeq>
    <INSDSeq_length>30</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..30</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>organism</INSDQualifier_name>
            <INSDQualifier_value>synthetic construct</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>mol_type</INSDQualifier_name>
            <INSDQualifier_value>other DNA</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>misc_feature</INSDFeature_key>
        <INSDFeature_location>1</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>This sequence is one of four branches of a branched
lynucleotide</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>cacacaaaaaaaaaaaaaaaaaaaaaaaaa</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
```

ST.26 paragraph 7 – sequences required in a SL

7. For the purpose of this Standard, a sequence for which inclusion in a sequence listing is required is one that is disclosed anywhere in an application by enumeration of its residues and can be represented as:

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

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

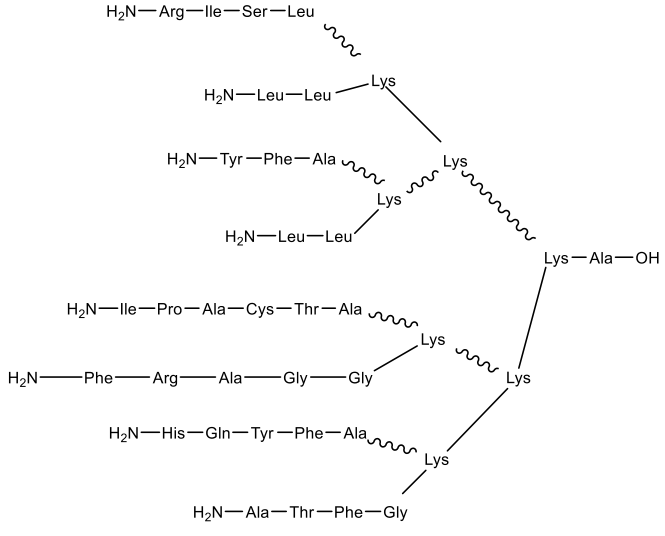
(ii) any chemical bond that results in an arrangement of adjacent nucleobases that mimics the arrangement of nucleobases in naturally occurring nucleic acids; or

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

ST.26 Annex VI – Example 7(b)-2 (1)

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

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



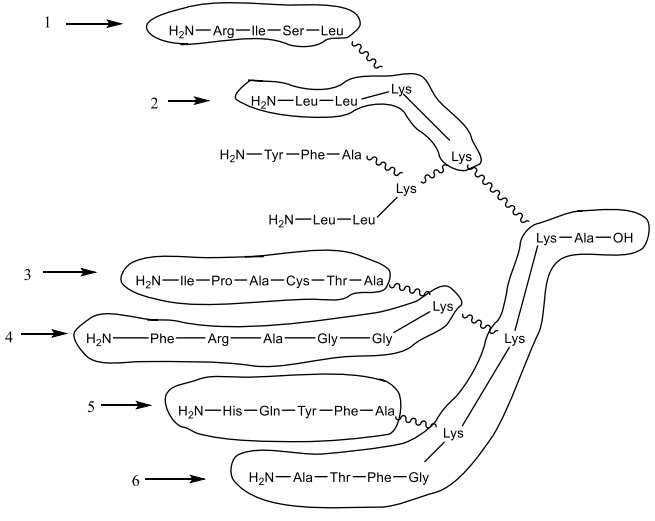
In the above branched peptide, the bonds between lysine and another amino acid depicted by ——— represent an amide linkage between the terminal amine of the lysine and the carboxyl end of the bonded amino acid. The bonds depicted by ∿∿∿ represent an amide linkage between the side chain amine of the lysine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

ST.26 Annex VI – Example 7(b)-2 (2)

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

YES

The example discloses a branched sequence where the lysine residues are used as a scaffolding. Paragraph 7(b) requires that the unbranched or linear region of the sequence, containing four or more specifically defined amino acids, be included in a sequence listing. In the above example, the linear regions of the branched peptide that have four or more specifically defined amino acids are encircled:



ST.26 paragraph 7(b) requires inclusion of peptides 1-6 above in a sequence listing.

Peptides which are not required, to be included in the sequence listing are:

YFA

LLK

ST.26 Annex VI – Example 7(b)-2 (3)

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

NO

According to paragraph 8, a sequence listing must not include any sequences having fewer than four specifically defined amino acids.

The peptides YFA and LLK each contain only three specifically defiend amino acids and therefore, they must not be included in the sequence listing as separate sequences with their own sequence identification numbers.

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

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

RISL (SEQ ID NO: 26)

LLKK (SEQ ID NO: 27)

IPACTA (SEQ ID NO: 28)

FRAGGK (SEQ ID NO: 29)

HQYFA (SEQ ID NO: 30)

ATFGKKKA (SEQ ID NO: 31)

The 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 amino acid sequence". According to ST.26 paragraph 29, SEQ ID Nos 27, 29, and 31, must include an annotation for each lysine to indicate that it is a modified amino acid, using the feature key "SITE" together with the qualifier "NOTE" describing that the side chain of the lysine is linked via an amide linkage to another sequence. Preferably, each of the SEQ ID Nos 26, 28, and 30 should include an annotation to indicate that the C-terminal amino acid is linked to another sequence, using the feature key "SITE" together with the qualifier "NOTE".

Relevant ST.26 paragraph(s): 7(b), 8, 26, 29, 30, and 31

ST.26 Annex VI Appendix SEQ ID NO: 27 (1)

```
<SequenceData sequenceIDNumber="27">
  <INSDSeq>
    <INSDSeq_length>4</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..4</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>synthetic construct</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>SITE</INSDFeature_key>
        <INSDFeature_location>1..4</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>This sequence is one part of a branched amino acid
sequence</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
```

ST.26 Annex VI Appendix SEQ ID NO: 27 (2)

```
            <INSDFeature>
              <INSDFeature_key>SITE</INSDFeature_key>
              <INSDFeature_location>3</INSDFeature_location>
              <INSDFeature_quals>
                <INSDQualifier>
                  <INSDQualifier_name>NOTE</INSDQualifier_name>
                  <INSDQualifier_value>The side chain of Lysine is linked via an amide linkage to another
sequence</INSDQualifier_value>
                </INSDQualifier>
              </INSDFeature_quals>
            </INSDFeature>
            <INSDFeature>
              <INSDFeature_key>SITE</INSDFeature_key>
              <INSDFeature_location>4</INSDFeature_location>
              <INSDFeature_quals>
                <INSDQualifier>
                  <INSDQualifier_name>NOTE</INSDQualifier_name>
                  <INSDQualifier_value>The side chain of Lysine is linked via an amide linkage to another
sequence</INSDQualifier_value>
                </INSDQualifier>
              </INSDFeature_quals>
            </INSDFeature>
          </INSDSeq_feature-table>
          <INSDSeq_sequence>LLKK</INSDSeq_sequence>
        </INSDSeq>
      </SequenceData>
```

ST.26 Annex VII – transformation of a SL from ST.25 to ST.26

- Requirements of ST.26 differ from ST.25
- Covers ST.26 mandatory requirements and any potential consequences of those requirements
- Transformation alone should not result in added or deleted matter, but caution needed
- Scenarios presented - 20
  - Explanations and recommendations provided

ST.26 Annex VII – Scenario 8 (1)

Scenario 8

ST.25 contains a number of feature keys that are not contained in ST.26. Therefore, applicants must take care to capture the information contained in those ST.25 feature keys in a manner compliant with ST.26 without the introduction of added or deleted subject matter.

Recommendations:

The following table provides guidance as to the manner in which the information contained in a former ST.25 feature key may be included in compliance with ST.26 without the introduction of added or deleted subject matter. Numbers 1-23 are feature keys related to nucleotide sequences and numbers 24 – 43 are feature keys related to amino acid sequences.

No.	ST.25 Feature key <221>	ST.26 equivalent		
		Feature key	Qualifier	Qualifier value
1	allele	misc_feature	allele	<223> value
2	attenuator	regulatory2	regulatory_class[2]	"attenuator"
			note *(if <223> present)*	<223> value
3	CAAT_signal	regulatory[2]	regulatory_class[2]	"CAAT_signal"
			note *(if <223> present)*	<223> value
4	conflict	misc_feature	note	"conflict" and <223> value
5	enhancer	regulatory[2]	regulatory_class[2]	"enhancer"
			note *(if <223> present)*	<223> value
6	GC_signal	regulatory[2]	regulatory_class[2]	"GC_signal"
			note *(if <223> present)*	<223> value
7	LTR	mobile_element[2]	rpt_type[2]	"long_terminal_repeat"
			note *(if <223> present)*	<223> value
8	misc_signal	regulatory[2]	regulatory_class[2]	"other"
			note *(if <223> present)*	<223> value
9	mutation	variation	note	"mutation" and <223> value
10	old_sequence	misc_feature	note	"old_sequence" and <223> value

ST.26 Annex VII – Scenario 8 (2)

No.	ST.25 Feature key <221>	ST.26 equivalent		
		Feature key	Qualifier	Qualifier value
24	NON_CONS	This feature relates to a gap of an unknown number of residues in a single sequence, which is prohibited in both ST.25 (paragraph 22) and ST.26 (paragraph 37). Consequently, each region of specifically defined residues that is encompassed by ST.26 paragraph 7 must be included in the sequence listing as a separate sequence and assigned its own sequence identification number. To avoid added/deleted subject matter, each such sequence must be annotated to indicate that it is part of a larger sequence that contains an undefined gap.		
		SITE	NOTE	Description
		Description - as to where and to what the sequence is linked, e.g. this residue is linked N-terminally to a peptide having an N-terminal Gly-Gly and a gap of undefined length.		
25	SIMILAR	REGION	NOTE	"SIMILAR" and <223> value if present
26	THIOETH	CROSSLNK	NOTE	"THIOETH" and <223> value if present
		For further location information guidance, see ST.26 Annex I, CROSSLNK Feature Key Comment		
27	THIOLEST	CROSSLNK	NOTE	"THIOLEST" and <223> value if present
		For further location information guidance, see ST.26 Annex I, CROSSLNK Feature Key Comment		
28	VARSPLIC	Discussed in a Scenario 13 below		
29	ACETYLATION	MOD_RES	NOTE	"ACETYLATION" and <223> value if present
			NOTE	Information required by ST.26 Annex I MOD_RES Feature Key Comment, if possible (without added subject matter)
30	AMIDATION	MOD_RES	NOTE	"AMIDATION" and <223> value if present
			NOTE	Information required by ST.26 Annex I MOD_RES Feature Key Comment, if possible (without added subject matter)
31	BLOCKED	MOD_RES	NOTE	"BLOCKED" and <223> value if present
			NOTE	Information required by ST.26 Annex I MOD_RES Feature Key Comment, if possible (without added subject matter)
32	FORMYLATION	MOD_RES	NOTE	"FORMYLATION" and <223> value if present
			NOTE	Information required by ST.26 Annex I MOD_RES Feature Key Comment, if possible (without added subject matter)

Questions?


UNITED STATES PATENT AND TRADEMARK OFFICE
DEPARTMENT OF COMMERCE