WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

WIPO Standard ST.26 ADVANCED

Webinar training

Note: this is a follow-on module and assumes that participants have already attended ST.26 Basics.

The ST.26 Basics course can be accessed here: https://www.wipo.int/meetings/en/details.jsp?meeting_id=62848

To be covered today

- Commonly used Feature keys and Qualifiers
- Feature Location Formats
- Qualifier Value Formats and Non-English Language Qualifier Values
- Special Situations - Uracil in DNA and Thymine in RNA; DNA/RNA hybrid molecules
- Nucleotide Analogs, D-amino acids, and Branched Sequences
- Sequence Variants

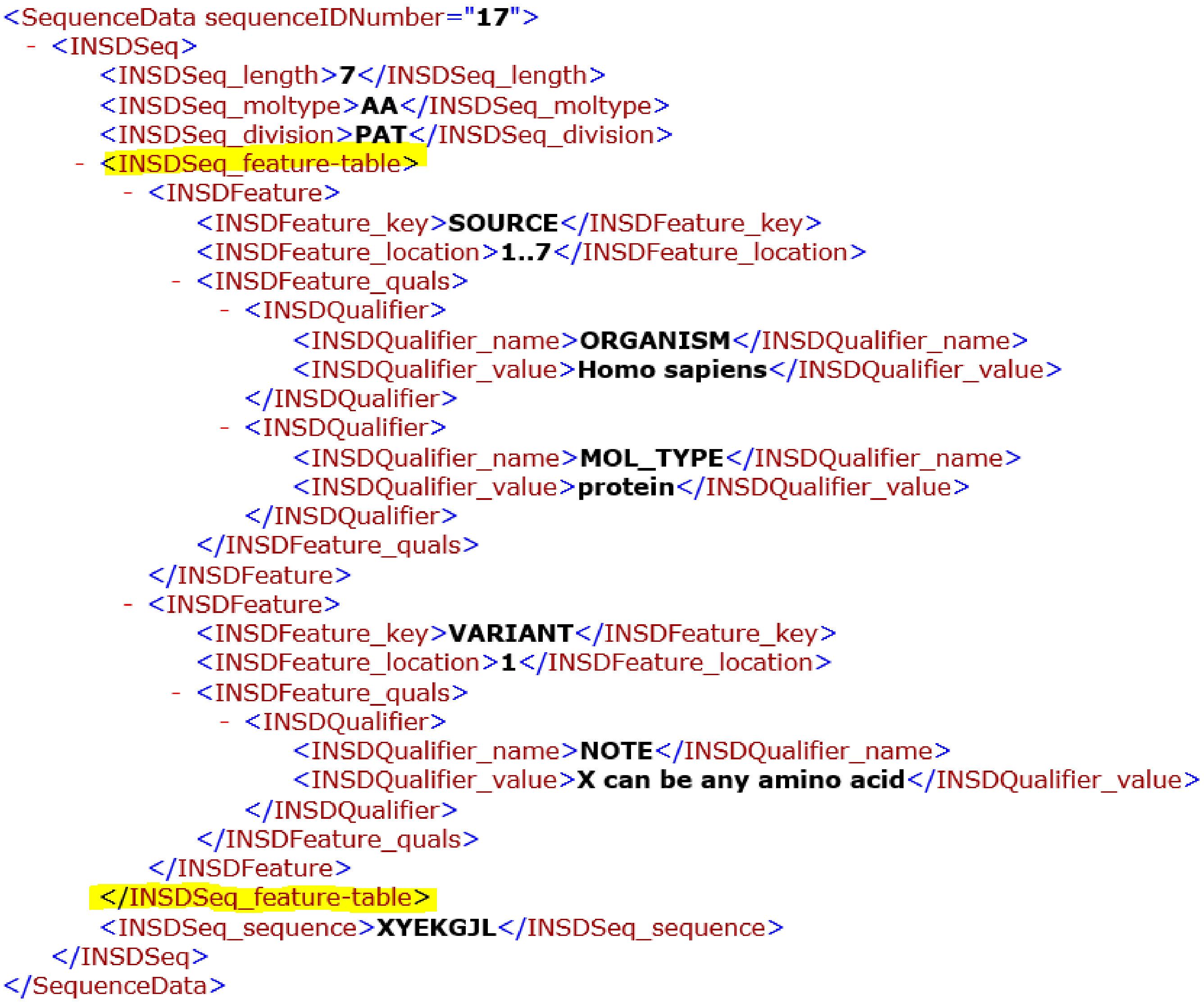
Commonly Used Feature Keys and Qualifiers

Feature Keys and Qualifiers

- Feature Keys can be used to further describe one or more residues of a sequence identified by location
  - Feature Keys for nucleotide sequences are listed in Annex I, Section 5
  - Feature keys for nucleotide sequences are lower-case; for example, “misc_binding”
  - Feature Keys for amino acid sequences are listed in Annex I, Section 7
  - Feature Keys for amino acid sequences are upper-case; for example, “REGION”

- Qualifiers can be used to further describe Features
  - Qualifiers for nucleotide sequences are listed in Annex I, Section 6
  - Qualifiers for nucleotide sequences are lower-case; for example, “allele”
  - Qualifiers for amino acid sequences are listed in Annex I, Section 8
  - Qualifiers for amino acid sequences are upper-case; for example, “NOTE”

Feature Keys and Qualifiers

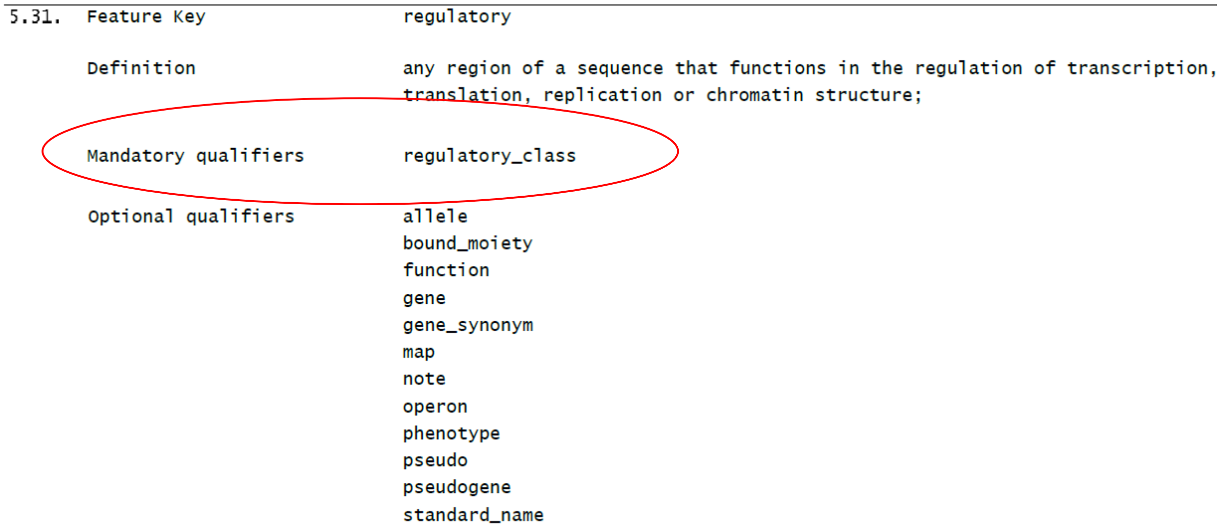


```
<SequenceData sequenceIDNumber="17">
 - <INSDSeq>
     <INSDSeq_length>7</INSDSeq_length>
     <INSDSeq_moltype>AA</INSDSeq_moltype>
     <INSDSeq_division>PAT</INSDSeq_division>
   - <INSDSeq_feature-table>
     - <INSDFeature>
         <INSDFeature_key>SOURCE</INSDFeature_key>
         <INSDFeature_location>1..7</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>
     - <INSDFeature>
         <INSDFeature_key>VARIANT</INSDFeature_key>
         <INSDFeature_location>1</INSDFeature_location>
       - <INSDFeature_quals>
         - <INSDQualifier>
             <INSDQualifier_name>NOTE</INSDQualifier_name>
             <INSDQualifier_value>X can be any amino acid</INSDQualifier_value>
           </INSDQualifier>
         </INSDFeature_quals>
       </INSDFeature>
     </INSDSeq_feature-table>
     <INSDSeq_sequence>XYEKGJL</INSDSeq_sequence>
   </INSDSeq>
</SequenceData>
```

Feature Keys and Qualifiers

- Feature Keys are optional except for the “source”/“SOURCE” feature, which is mandatory for every sequence.

- Each Feature Key will have a list of qualifiers that may be used to further describe the Feature. Most qualifiers are optional; however, some Feature Keys have mandatory qualifiers.

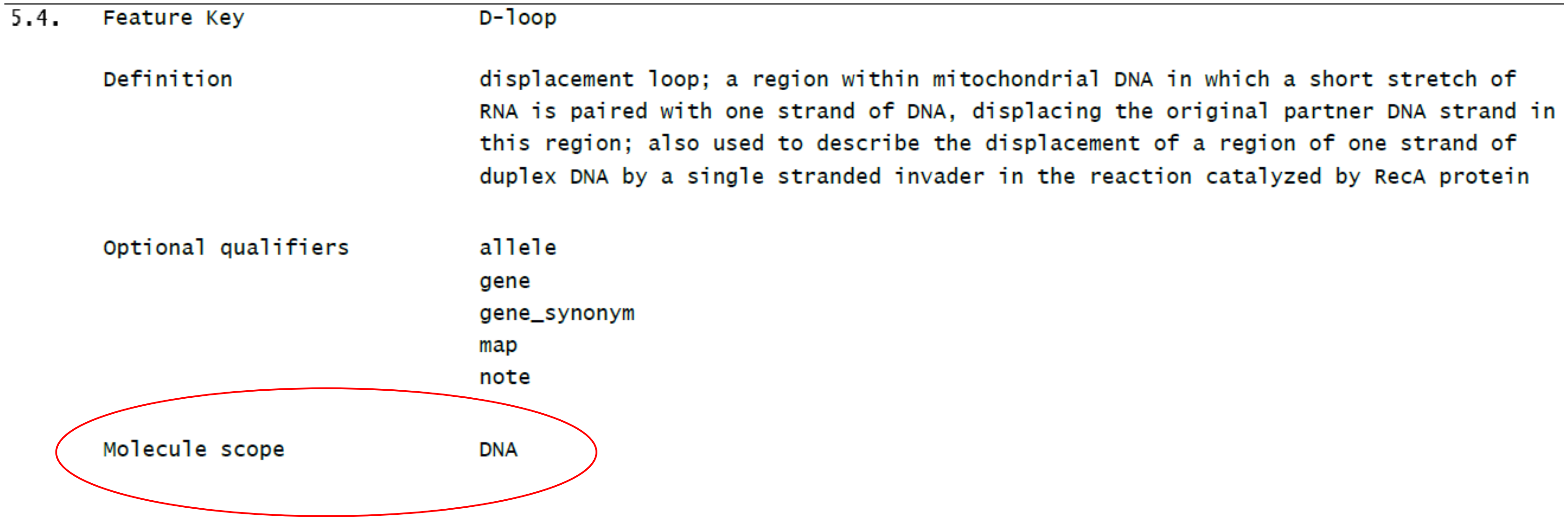


5.31.	Feature Key	regulatory
	Definition	any region of a sequence that functions in the regulation of transcription, translation, replication or chromatin structure;
	Mandatory qualifiers	regulatory_class
	Optional qualifiers	allele bound_moiety function gene gene_synonym map note operon phenotype pseudo pseudogene standard_name

- Qualifiers “mol_type”/“MOL_TYPE” and “organism”/“ORGANISM” are mandatory for the “source”/“SOURCE” feature.

Feature Keys and Qualifiers

- Some Feature Keys have additional limitations
  - organism scope; for example, “C_region” is limited to eukaryotes
  - molecule scope; for example, “D-loop” is limited to DNA sequences



5.4.	Feature Key	D-loop
	Definition	displacement loop; a region within mitochondrial DNA in which a short stretch of RNA is paired with one strand of DNA, displacing the original partner DNA strand in this region; also used to describe the displacement of a region of one strand of duplex DNA by a single stranded invader in the reaction catalyzed by RecA protein
	Optional qualifiers	allele gene gene_synonym map note
	Molecule scope	DNA

Feature Keys and Qualifiers

Nucleotide sequences: “modified_base” Feature Key

- The Feature Key “modified_base” and its mandatory qualifier “mod_base” should be used to describe a modified nucleotide (ST.26, paragraph 16)

- A “modified nucleotide” is any nucleotide other than:
  - deoxy-[a, g, c, or t] 3’-monophosphate
  - [a, g, c, or u] 3’-monophosphate (ST.26, paragraph 3(f))

- A “modified nucleotide” should be represented by the corresponding unmodified nucleotide (in Annex 1, Section 1, Table 1), whenever possible. Otherwise, it can be represented by “n”. For example, “2’-O-methylcytidine” should be represented by “c” in the sequence. “Queuosine” should be represented by “n”. The symbol “n” is equivalent to only one residue.

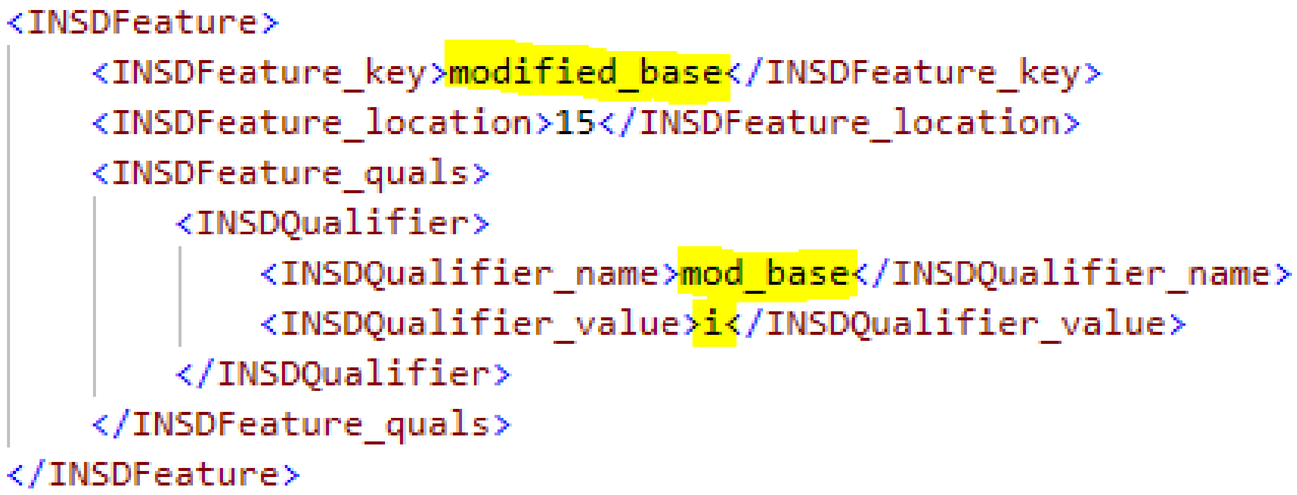
- The value of the mandatory qualifier “mod_base” must be selected from the values in Annex I, Section 2, Table 2. If “other” is the value, then an additional “note” qualifier must contain the complete, unabbreviated name of the modified residue.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Feature Keys and Qualifiers

Nucleotide sequences: “modified_base” Feature Key

- Example: nucleotide sequence with inosine at position 15



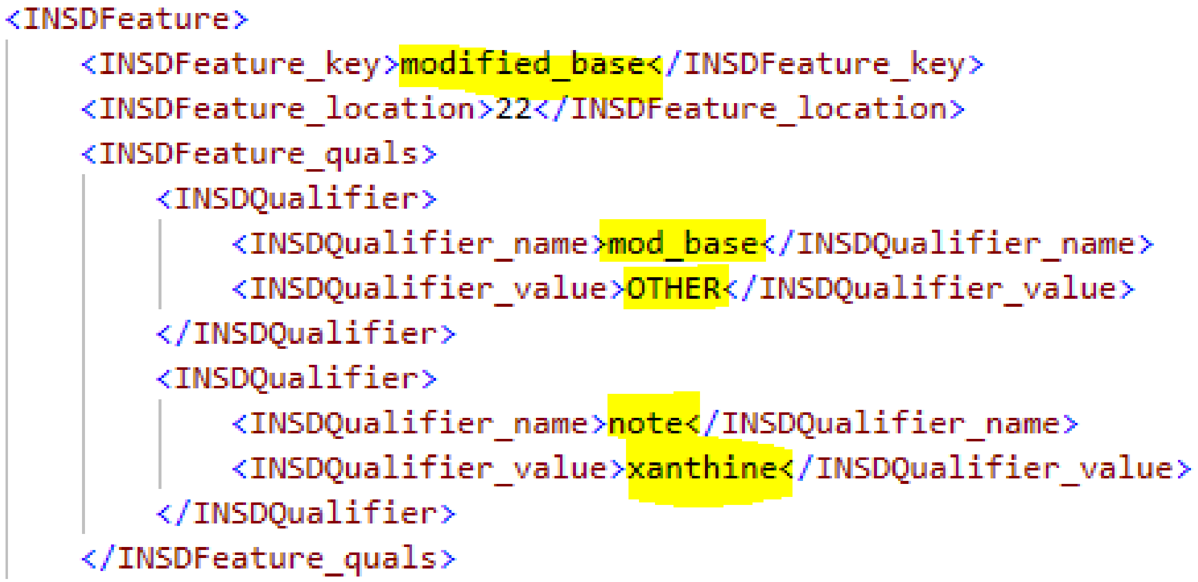
```xml
<INSDFeature>
    <INSDFeature_key>modified_base</INSDFeature_key>
    <INSDFeature_location>15</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>mod_base</INSDQualifier_name>
            <INSDQualifier_value>i</INSDQualifier_value>
       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

- “Inosine” is listed in Annex I, Section 2, Table 2 with the abbreviation “i”

Feature Keys and Qualifiers

Nucleotide sequences: “modified_base” Feature Key

- Example: nucleotide sequence with xanthine at position 22



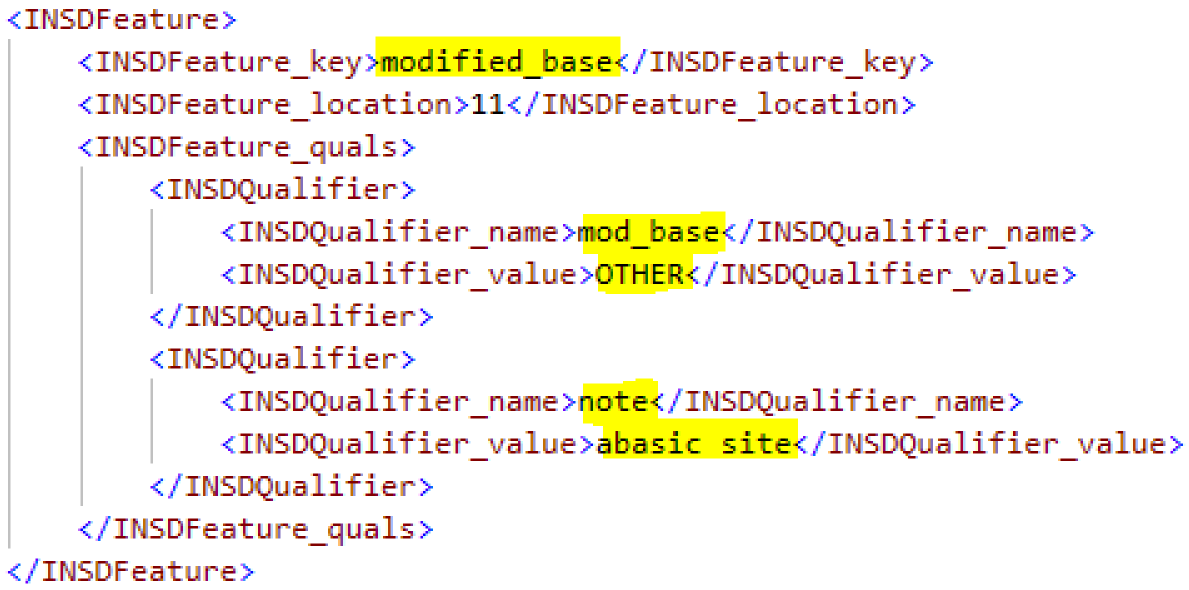
```
<INSDFeature>
    <INSDFeature_key>modified_base</INSDFeature_key>
    <INSDFeature_location>22</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>mod_base</INSDQualifier_name>
            <INSDQualifier_value>OTHER</INSDQualifier_value>
        </INSDQualifier>
        <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>xanthine</INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
```

- “Xanthine” is NOT listed in Annex I, Section 2, Table 2; therefore, the value for the mandatory “mod_base” qualifier must be “OTHER” and an additional “note” qualifier must be included with the value “xanthine.”

Feature Keys and Qualifiers

Nucleotide sequences: “modified_base” Feature Key

- “modified_base” can also be used to describe an abasic site:



```
<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>
```

- The abasic site can be represented by an “n” in the sequence, and further described using a “modified_base” feature key along with a “mod_base” qualifier with the value “OTHER” and an additional “note” qualifier with the value “abasic site”.

Feature Keys and Qualifiers

Nucleotide sequences: “CDS” Feature Key

- The Feature Key “CDS” may be used to identify coding sequences. The location of the CDS feature must include the stop codon. (ST.26, pgh. 89)

- There are NO mandatory qualifiers for the “CDS” Feature Key

- Commonly used qualifiers include:

“pseudo”

“transl_table”

“pseudogene”

“codon_start”

“protein_id”

“translation”

“transl_except”

Feature Keys and Qualifiers

Nucleotide sequences: “CDS” Feature Key

- A “CDS” feature can have only ONE of the following qualifiers:

Qualifier	Description	Value
pseudo	indicates that the CDS feature is non-functional and has no translation, but is not a pseudogene	none
pseudogene	indicates that the CDS feature is a pseudogene and has no translation	processed unprocessed unitary allelic unknown
translation	indicates the amino acid sequence derived from translating the CDS	one-letter amino acid abbreviations

Feature Keys and Qualifiers

Nucleotide sequences: “CDS” Feature Key

- An amino acid sequence containing 4 or more specifically defined amino acids that is encoded by a coding sequence and disclosed in a “translation” qualifier must be included in the sequence listing as a separate sequence and assigned its own sequence identification number.

- The sequence identification number assigned to the amino acid sequence must be provided as the value in the qualifier “protein_id” within the “CDS” feature key.

- The “ORGANISM” qualifier of the “SOURCE” feature key for the amino acid sequence must be identical to that of its coding sequence. (ST.26, pgh. 92)

Feature Keys and Qualifiers

Nucleotide sequences: “CDS” Feature Key

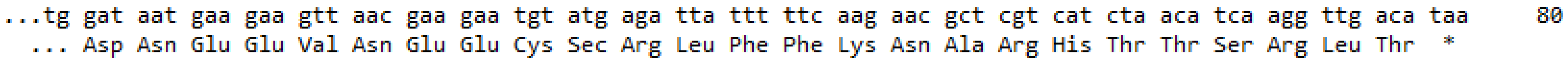
- Qualifiers that can modify the translated sequence:

Qualifier	Description	Value
transl_table	indicates the genetic code table used to translate the CDS; default is “1-Standard Code”	a number that corresponds to a translation table in Annex I, Section 9
transl_except	indicates the translation of a codon that does not conform to the genetic code defined in “transl_table”	(pos:<location>, aa:<amino_acid>)
codon_start	indicates the reading frame of the CDS relative to the first base	1, 2, or 3

Feature Keys and Qualifiers

Nucleotide sequences: “CDS” Feature Key

- Example - SEQ ID NO:1 is a fragment of a coding sequence from a yeast mitochondrial gene (genetic code table “3-Yeast Mitochondrial Code”):



```
...tg gat aat gaa gaa gtt aac gaa gaa tgt atg aga tta ttt ttc aag aac gct cgt cat cta aca tca agg ttg aca taa     80
  ... Asp Asn Glu Glu Val Asn Glu Glu Cys Sec Arg Leu Phe Phe Lys Asn Ala Arg His Thr Thr Ser Arg Leu Thr  *
```

The codon at positions 30-32 encode selenocysteine (Sec)

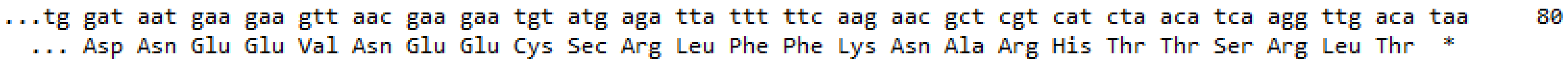
The sequence begins with a partial codon

- What information should be included in a CDS feature to accurately represent this sequence?

Feature Keys and Qualifiers

Nucleotide sequences: “CDS” Feature Key

- Example - SEQ ID NO:1 is a fragment of a coding sequence from a yeast mitochondrial gene (genetic code table “3-Yeast Mitochondrial Code”):



```
...tg gat aat gaa gaa gtt aac gaa gaa tgt atg aga tta ttt ttc aag aac gct cgt cat cta aca tca agg ttg aca taa     80
  ... Asp Asn Glu Glu Val Asn Glu Glu Cys Sec Arg Leu Phe Phe Lys Asn Ala Arg His Thr Thr Ser Arg Leu Thr  *
```

- Sequence represented in the sequence listing as:

tggataatga agaagttaac gaagaatgta tgagattatt tttcaagaac gctcgtcatc taacatcaag gttgacataa

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Feature Keys and Qualifiers

Nucleotide sequences: “CDS” Feature Key

- Example - SEQ ID NO:1 is a fragment of a coding sequence from a yeast mitochondrial gene (genetic code table “3-Yeast Mitochondrial Code”):

```
...tg gat aat gaa gaa gtt aac gaa gaa tgt atg aga tta ttt ttc aag aac gct cgt cat cta aca tca agg ttg aca taa     80
  ... Asp Asn Glu Glu Val Asn Glu Glu Cys Sec Arg Leu Phe Phe Lys Asn Ala Arg His Thr Thr Ser Arg Leu Thr  *
```

- “CDS” Feature Key
- Feature location: <1..80

The location includes the stop codon at positions 78-80

The “<“ indicates that the coding region begins before position 1

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Feature Keys and Qualifiers

Nucleotide sequences: “CDS” Feature Key

- Example - SEQ ID NO:1 is a fragment of a coding sequence from a yeast mitochondrial gene (genetic code table “3-Yeast Mitochondrial Code”):

```
...tg gat aat gaa gaa gtt aac gaa gaa tgt atg aga tta ttt ttc aag aac gct cgt cat cta aca tca agg ttg aca taa     80
  ... Asp Asn Glu Glu Val Asn Glu Glu Cys Sec Arg Leu Phe Phe Lys Asn Ala Arg His Thr Thr Ser Arg Leu Thr  *
```

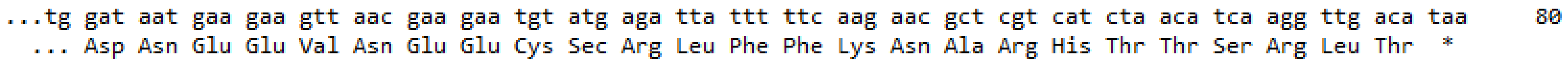
- “CDS” Feature Key
- Feature location: <1..80
- “codon_start” qualifier with value “3”

The “codon_start” qualifier value of 3 indicates that the first full codon begins at the third position within the location

Feature Keys and Qualifiers

Nucleotide sequences: “CDS” Feature Key

- Example - SEQ ID NO:1 is a fragment of a coding sequence from a yeast mitochondrial gene (genetic code table “3-Yeast Mitochondrial Code”):



```
...tg gat aat gaa gaa gtt aac gaa gaa tgt atg aga tta ttt ttc aag aac gct cgt cat cta aca tca agg ttg aca taa     80
  ... Asp Asn Glu Glu Val Asn Glu Glu Cys Sec Arg Leu Phe Phe Lys Asn Ala Arg His Thr Thr Ser Arg Leu Thr  *
```

- “CDS” Feature Key
- Feature location: <1..80
- “codon_start” qualifier with value “3”
- “transl_table” qualifier with the value “3”

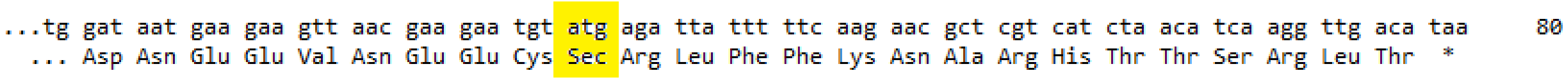
The Genetic Code Tables in Annex 1, Section 9, are used determine the value of the “transl_table” qualifier.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Feature Keys and Qualifiers

Nucleotide sequences: “CDS” Feature Key

- Example - SEQ ID NO:1 is a fragment of a coding sequence from a yeast mitochondrial gene (genetic code table “3-Yeast Mitochondrial Code”):



```
...tg gat aat gaa gaa gtt aac gaa gaa tgt atg aga tta ttt ttc aag aac gct cgt cat cta aca tca agg ttg aca taa     80
  ... Asp Asn Glu Glu Val Asn Glu Glu Cys Sec Arg Leu Phe Phe Lys Asn Ala Arg His Thr Thr Ser Arg Leu Thr  *
```

- “CDS” Feature Key
- Feature location: <1..80
- “codon_start” qualifier with value “3”
- “transl_table” qualifier with the value “3”
- “transl_except” qualifier with the value “(pos:30..32,aa:Sec)”

The “transl_except” qualifier will put in the amino acid selenocysteine

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Feature Keys and Qualifiers

Nucleotide sequences: “CDS” Feature Key

- Example - SEQ ID NO:1 is a fragment of a coding sequence from a yeast mitochondrial gene (genetic code table “3-Yeast Mitochondrial Code”):

```
...tg gat aat gaa gaa gtt aac gaa gaa tgt atg aga tta ttt ttc aag aac gct cgt cat cta aca tca agg ttg aca taa      80
  ... Asp Asn Glu Glu Val Asn Glu Glu Cys Sec Arg Leu Phe Phe Lys Asn Ala Arg His Thr Thr Ser Arg Leu Thr  *
```

- “CDS” Feature Key
- Feature location: <1..80
- “codon_start” qualifier with value “3”
- “transl_table” qualifier with the value “3”
- “transl_except” qualifier with the value “(pos:30..32,aa:Sec)”
- “translation” qualifier with the value “**DNEEVNEECURLFFKNARHTTSRLT**”

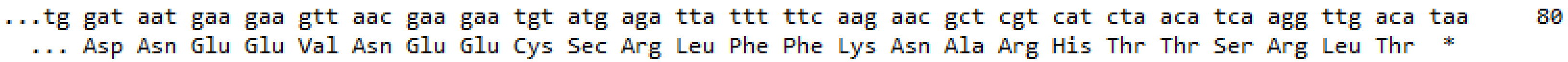
The stop codon is not shown in the translation qualifier!

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Feature Keys and Qualifiers

Nucleotide sequences: “CDS” Feature Key

- Example - SEQ ID NO:1 is a fragment of a coding sequence from a yeast mitochondrial gene (genetic code table “3-Yeast Mitochondrial Code”):



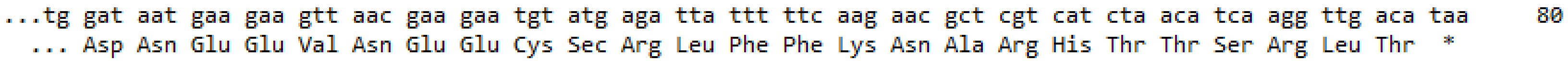
```
...tg gat aat gaa gaa gtt aac gaa gaa tgt atg aga tta ttt ttc aag aac gct cgt cat cta aca tca agg ttg aca taa     80
  ... Asp Asn Glu Glu Val Asn Glu Glu Cys Sec Arg Leu Phe Phe Lys Asn Ala Arg His Thr Thr Ser Arg Leu Thr  *
```

- “CDS” Feature Key
- Feature location: <1..80
- “codon_start” qualifier with value “3”
- “transl_table” qualifier with the value “3”
- “transl_except” qualifier with the value “(pos:30..32,aa:Sec)”
- “translation” qualifier with the value “**DNEEVNEECURLFFKNARHTTSRLT**”
- a separate protein sequence for the translation

Feature Keys and Qualifiers

Nucleotide sequences: “CDS” Feature Key

- Example - SEQ ID NO:1 is a fragment of a coding sequence from a yeast mitochondrial gene (genetic code table “3-Yeast Mitochondrial Code”):



```
...tg gat aat gaa gaa gtt aac gaa gaa tgt atg aga tta ttt ttc aag aac gct cgt cat cta aca tca agg ttg aca taa     80
  ... Asp Asn Glu Glu Val Asn Glu Glu Cys Sec Arg Leu Phe Phe Lys Asn Ala Arg His Thr Thr Ser Arg Leu Thr  *
```

- “CDS” Feature Key
- Feature location: <1..80
- “codon_start” qualifier with value “3”
- “transl_table” qualifier with the value “3”
- “transl_except” qualifier with the value “(pos:30..32,aa:Sec)”
- “translation” qualifier with the value “**DNEEVNEECURLFFKNARHTTSRLT**”
- a separate protein sequence for the translation
- “protein_id” qualifier with the SEQ ID number of the translated protein

Feature Keys and Qualifiers

Nucleotide sequences: “CDS” Feature Key

- The “CDS” feature location can use the “join” location operator to connect discontinuous segments of a sequence into a single coding region

join(location1,location2)

- The “CDS” feature location can use the “complement” operator to indicate that the feature is located on the strand complementary to the sequence specified by the location descriptor

complement(location)

Feature Keys and Qualifiers

Amino acid sequences: Commonly Used Feature Keys

Feature Key	Description	Mandatory Qualifiers?
SITE	indicates an interesting single amino-acid site	a mandatory NOTE qualifier must describe the site
REGION	indicates a region of interest	none; NOTE is optional
BINDING	indicates the binding site for a chemical group	a mandatory NOTE qualifier must contain the name of the chemical group
UNSURE	describes regions of uncertainty in the sequence	none; NOTE is optional

Feature Keys and Qualifiers

Amino acid sequences: Modified amino acids

- A “modified amino acid” is any amino acid other than:

  L-alanine
  L-arginine
  L-asparagine
  L-aspartic acid
  L-cysteine
  L-glutamine
  L-glutamic acid
  L-glycine
  L-histidine
  L-isoleucine
  L-leucine
  L-lysine
  L-methionine
  L-phenylalanine
  L-proline
  L-pyrrolysine
  L-serine
  L-selenocysteine
  L-threonine
  L-tryptophan
  L-tyrosine
  L-valine

  (ST.26, pgh. 3(e))

- A “modified amino acid” should be represented by the corresponding unmodified amino acid, whenever possible. Otherwise, it can be represented by “X”. For example, “hydroxylysine” should be represented by “K” in the sequence. “Ornithine” should be represented by “X”. (ST.26, pgh. 29).

- The symbol “X” is equivalent to only one residue.

Feature Keys and Qualifiers

Amino acid sequences: Modified amino acids

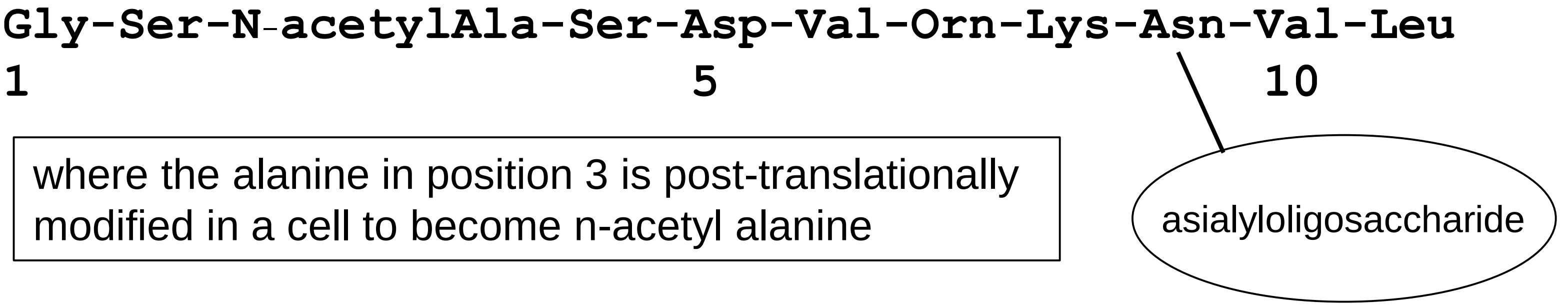
- Several Feature Keys can be used to indicate a “modified amino acid”:

Feature Key	Description
SITE	indicates a non post-translationally modified amino acid
MOD_RES	indicates a post-translationally modified amino acid
CARBOHYD	indicates a glycosylated amino acid
LIPID	indicates the covalent binding of a lipid moiety to an amino acid

- A mandatory NOTE qualifier must be included with each of the above feature keys, with a value that describes the modification.

Feature Keys and Qualifiers

Amino acid sequences: Modified amino acids



Gly-Ser-N-acetylAla-Ser-Asp-Val-Orn-Lys-Asn-Val-Leu

1 5 10

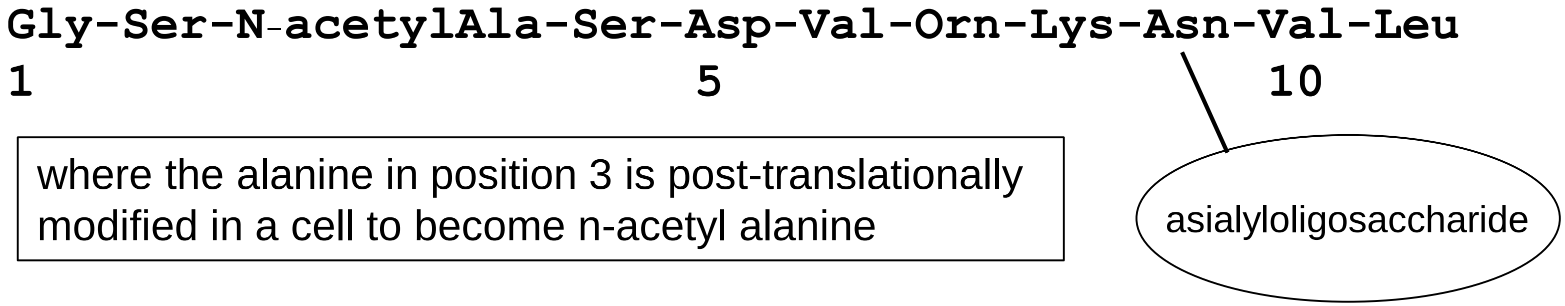
where the alanine in position 3 is post-translationally modified in a cell to become n-acetyl alanine

asialyloligosaccharide

- What Feature Keys and Qualifiers should be included in the sequence listing to accurately represent this sequence?

Feature Keys and Qualifiers

Amino acid sequences: Modified amino acids



Gly-Ser-N-acetylAla-Ser-Asp-Val-Orn-Lys-Asn-Val-Leu

1 5 10

where the alanine in position 3 is post-translationally modified in a cell to become n-acetyl alanine

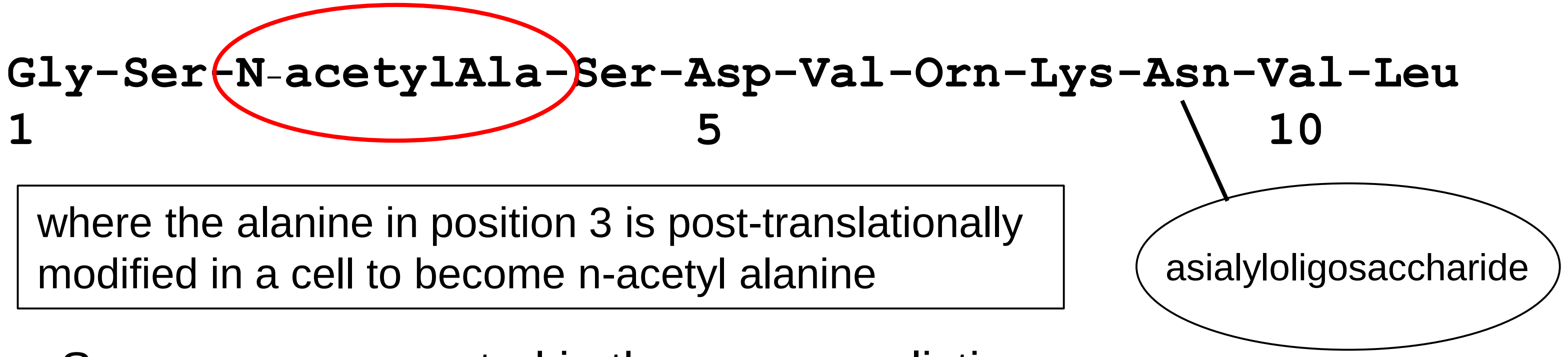
asialyloligosaccharide

- Sequence represented in the sequence listing as:

GSASDVXKNVL

Feature Keys and Qualifiers

Amino acid sequences: Modified amino acids



```
Gly-Ser-N-acetylAla-Ser-Asp-Val-Orn-Lys-Asn-Val-Leu
1                        5                    10
```

where the alanine in position 3 is post-translationally modified in a cell to become n-acetyl alanine

asialyloligosaccharide

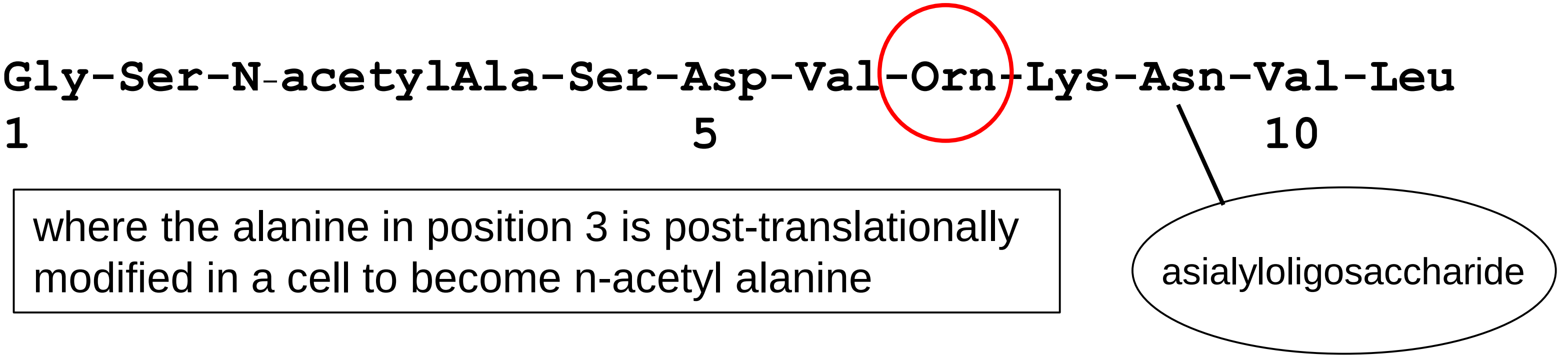
- Sequence represented in the sequence listing as:

  GSASDVXKNVL

- “MOD_RES” Feature Key with location “3” and NOTE qualifier with value “N-acetylalanine”

Feature Keys and Qualifiers

Amino acid sequences: Modified amino acids



```
Gly-Ser-N-acetylAla-Ser-Asp-Val-Orn-Lys-Asn-Val-Leu
1                         5                     10
```

where the alanine in position 3 is post-translationally modified in a cell to become n-acetyl alanine

asialyloligosaccharide

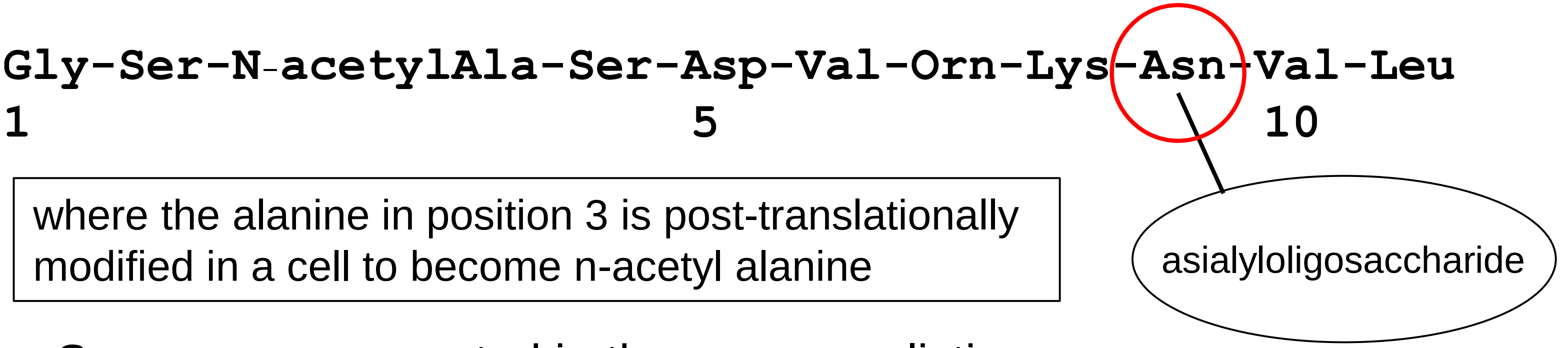
- Sequence represented in the sequence listing as:

  GSASDVXKNVL

- “MOD_RES” Feature Key with location “3” and NOTE qualifier with value “N-acetylalanine”
- “SITE” Feature Key with location “7” and NOTE qualifier with value “ornithine”

Feature Keys and Qualifiers

Amino acid sequences: Modified amino acids



Gly-Ser-N-acetylAla-Ser-Asp-Val-Orn-Lys-Asn-Val-Leu

1 5 10

where the alanine in position 3 is post-translationally modified in a cell to become n-acetyl alanine

- Sequence represented in the sequence listing as:

  GSASDVXKNVL

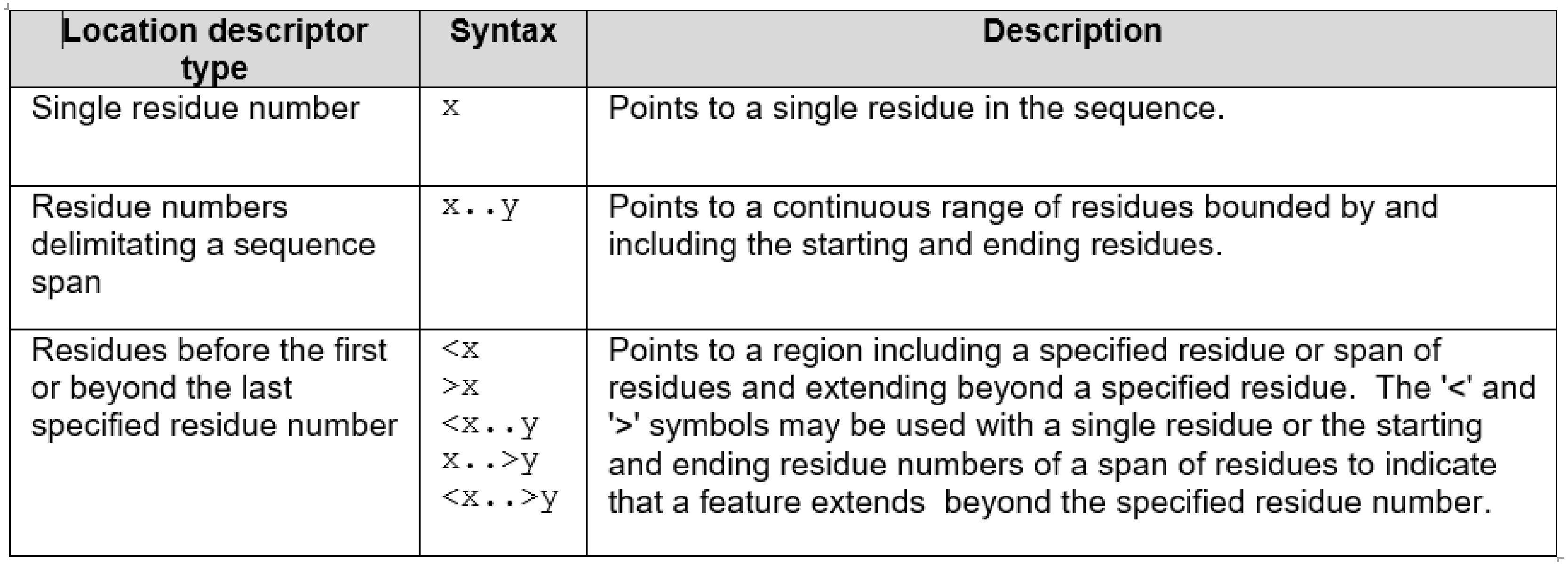
- “MOD_RES” Feature Key with location “3” and NOTE qualifier with value “N-acetylalanine”
- “SITE” Feature Key with location “7” and NOTE qualifier with value “ornithine”
- “CARBOHYD” Feature Key with location “9” and NOTE qualifier with value “Asn side-chain linked to asialyloligosaccharide”

Feature Location Formats

Location Formats

Location descriptors for all molecule types

- Location descriptors are used to identify the location of a feature in a sequence
- ST.26 has mandatory requirements for the format of location descriptors
- The following location descriptor formats can be used for nucleotide and amino acid sequences:



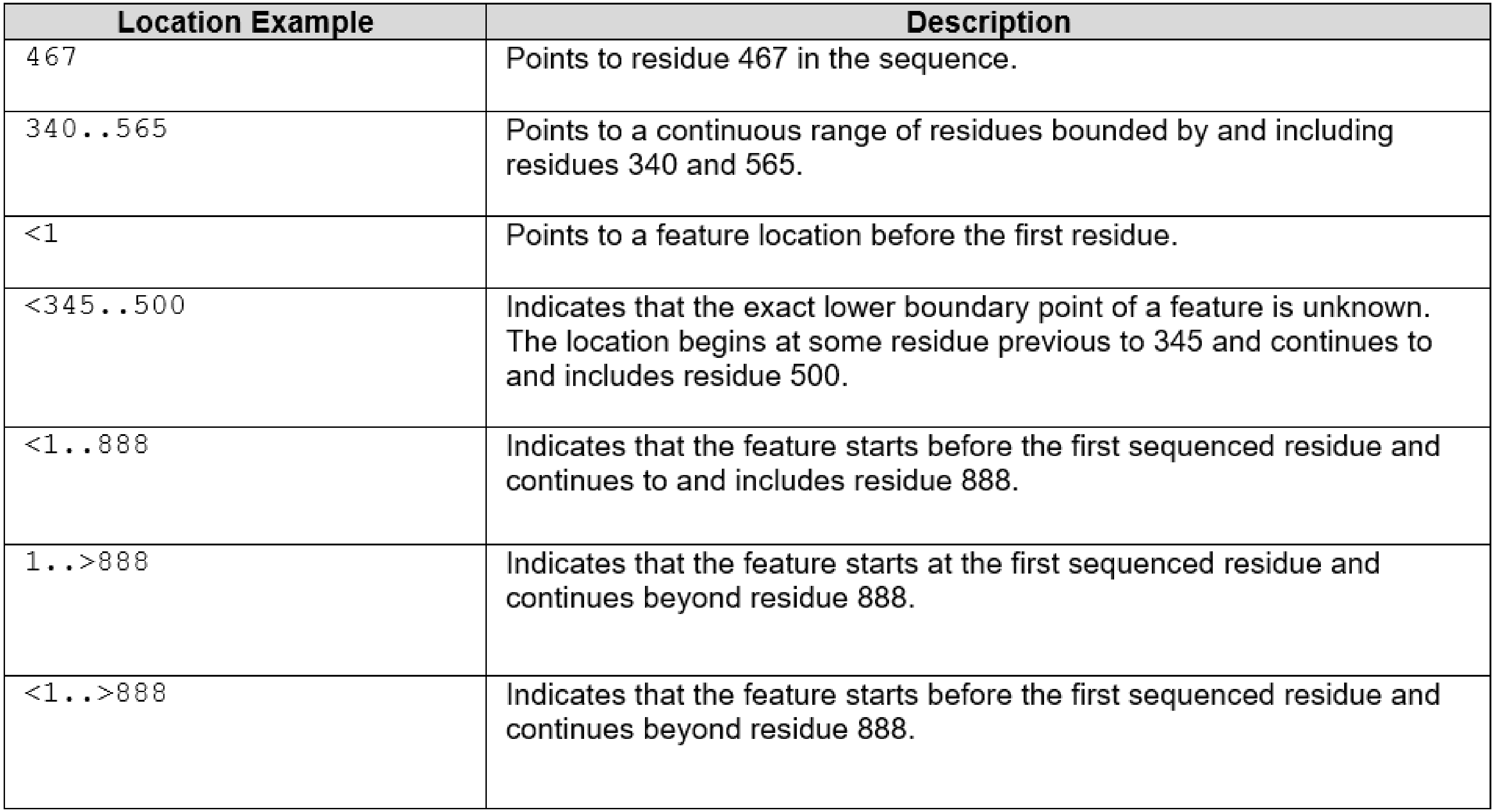
Location descriptor type	Syntax	Description
Single residue number	`x`	Points to a single residue in the sequence.
Residue numbers delimitating a sequence span	`x..y`	Points to a continuous range of residues bounded by and including the starting and ending residues.

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

WIPO Standard ST.26, paragraph 66(a)

Location Formats

Location descriptors for all molecule types



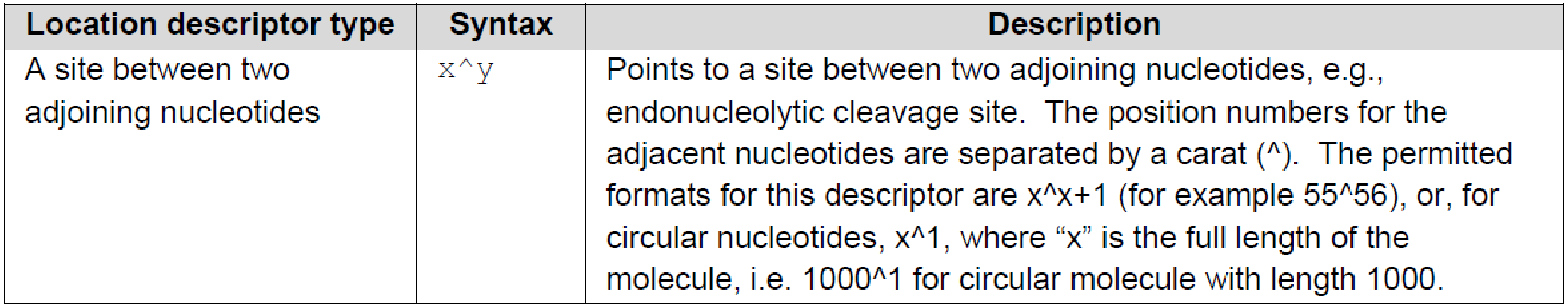
Location Example	Description
467	Points to residue 467 in the sequence.
340..565	Points to a continuous range of residues bounded by and including residues 340 and 565.
<1	Points to a feature location before the first residue.
<345..500	Indicates that the exact lower boundary point of a feature is unknown. The location begins at some residue previous to 345 and continues to and includes residue 500.
<1..888	Indicates that the feature starts before the first sequenced residue and continues to and includes residue 888.
1..>888	Indicates that the feature starts at the first sequenced residue and continues beyond residue 888.
<1..>888	Indicates that the feature starts before the first sequenced residue and continues beyond residue 888.

WIPO Standard ST.26, paragraph 70(a)

Location Formats

Location descriptors exclusively for nucleotide sequences

- The following location descriptor format can ONLY be used for DNA and RNA sequences:



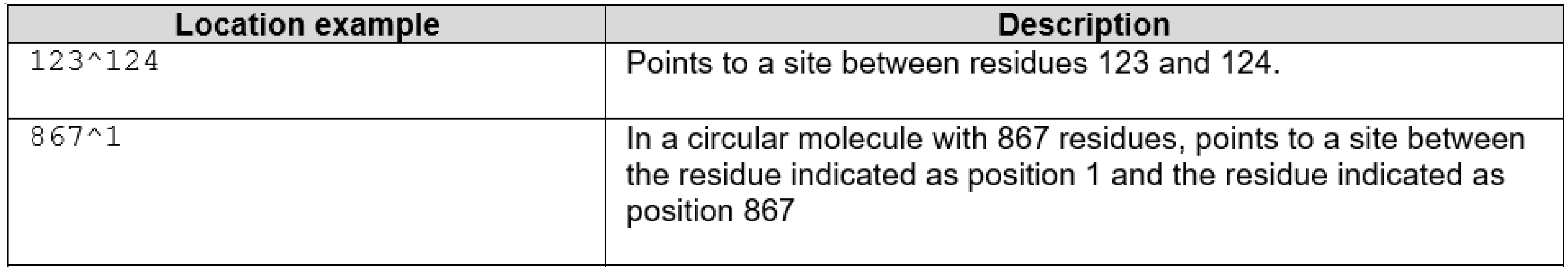
Location descriptor type	Syntax	Description
A site between two adjoining nucleotides	`x^y`	Points to a site between two adjoining nucleotides, e.g., endonucleolytic cleavage site. The position numbers for the adjacent nucleotides are separated by a carat (^). The permitted formats for this descriptor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

WIPO Standard ST.26, paragraph 66(b)

Location Formats

Location descriptors exclusively for nucleotide sequences

- When using the x^y location format, x and y must be adjacent residues



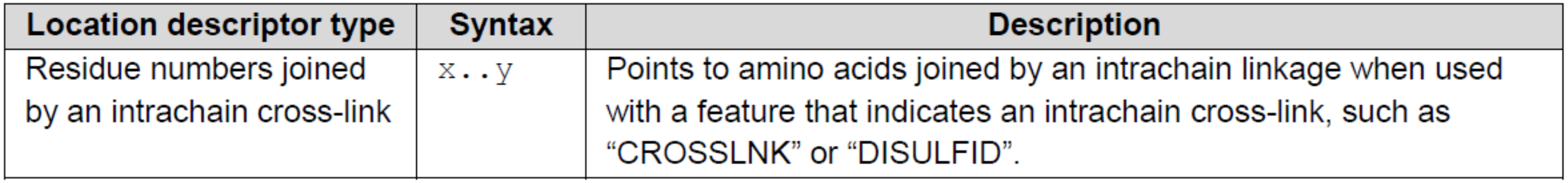
Location example	Description
`123^124`	Points to a site between residues 123 and 124.
`867^1`	In a circular molecule with 867 residues, points to a site between the residue indicated as position 1 and the residue indicated as position 867

WIPO Standard ST.26, paragraph 70(b)

Location Formats

Location descriptors – special case for amino acid sequences

- The location descriptor x..y indicates an intrachain crosslink between the two indicated residues when used with a “CROSSLNK” or “DISULFID” feature key

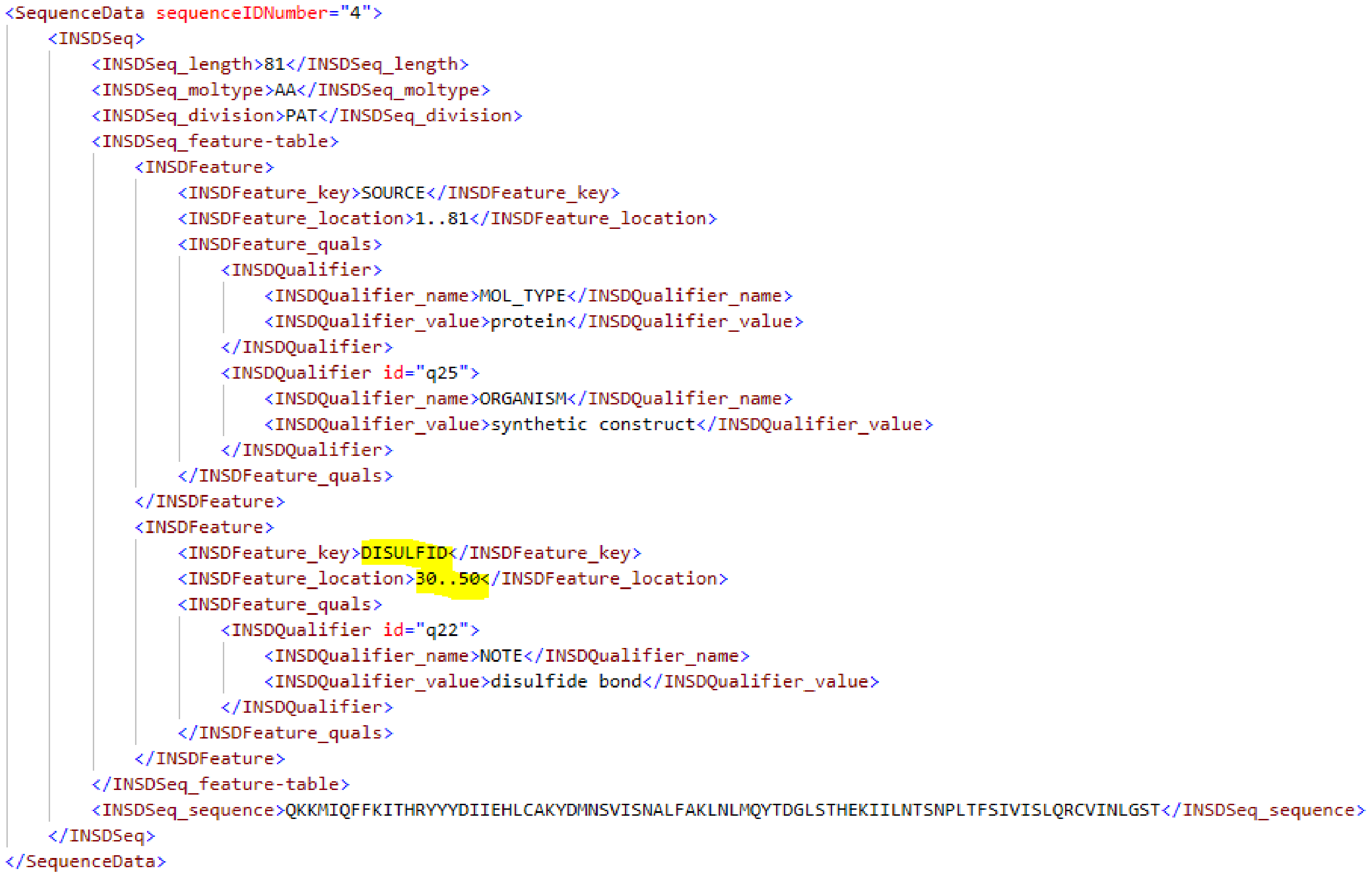


Location descriptor type	Syntax	Description
Residue numbers joined by an intrachain cross-link	`x..y`	Points to amino acids joined by an intrachain linkage when used with a feature that indicates an intrachain cross-link, such as “CROSSLNK” or “DISULFID”.

WIPO Standard ST.26, paragraph 66(c)

Location Formats

Location descriptors – special case for amino acid sequences

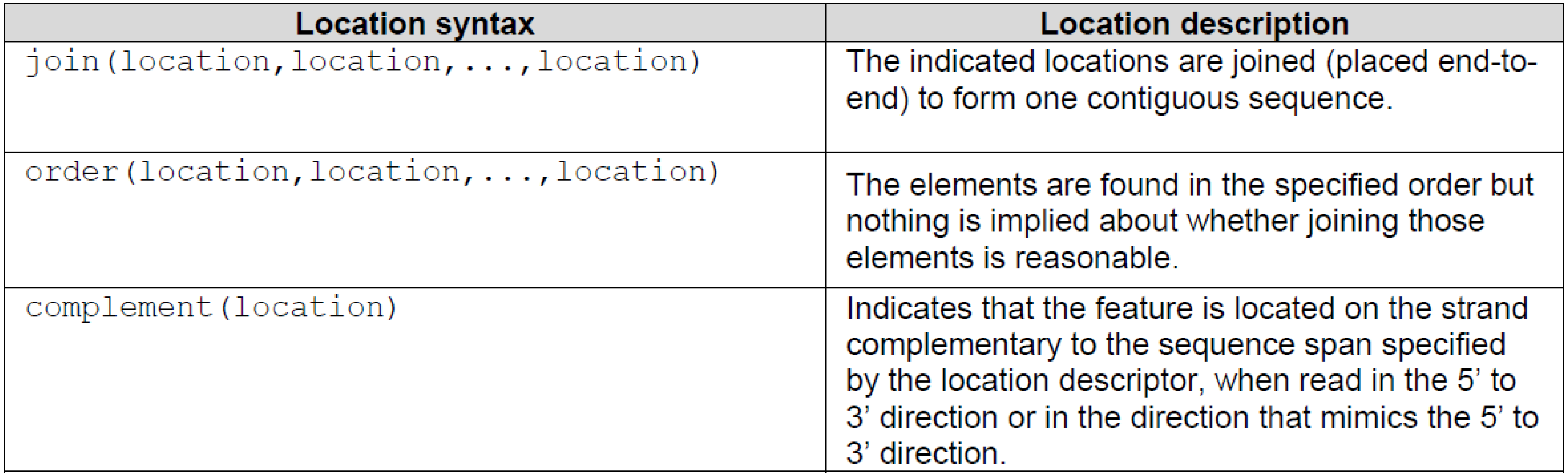


```
<SequenceData sequenceIDNumber="4">
    <INSDSeq>
        <INSDSeq_length>81</INSDSeq_length>
        <INSDSeq_moltype>AA</INSDSeq_moltype>
        <INSDSeq_division>PAT</INSDSeq_division>
        <INSDSeq_feature-table>
            <INSDFeature>
                <INSDFeature_key>SOURCE</INSDFeature_key>
                <INSDFeature_location>1..81</INSDFeature_location>
                <INSDFeature_quals>
                    <INSDQualifier>
                        <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
                        <INSDQualifier_value>protein</INSDQualifier_value>
                    </INSDQualifier>
                    <INSDQualifier id="q25">
                        <INSDQualifier_name>ORGANISM</INSDQualifier_name>
                        <INSDQualifier_value>synthetic construct</INSDQualifier_value>
                    </INSDQualifier>
                </INSDFeature_quals>
            </INSDFeature>
            <INSDFeature>
                <INSDFeature_key>DISULFID</INSDFeature_key>
                <INSDFeature_location>30..50</INSDFeature_location>
                <INSDFeature_quals>
                    <INSDQualifier id="q22">
                        <INSDQualifier_name>NOTE</INSDQualifier_name>
                        <INSDQualifier_value>disulfide bond</INSDQualifier_value>
                    </INSDQualifier>
                </INSDFeature_quals>
            </INSDFeature>
        </INSDSeq_feature-table>
        <INSDSeq_sequence>QKKMIQFFKITHRYYYDIIEHLCAKYDMNSVISNALFAKLNLMQYTDGLSTHEKIILNTSNPLTFSIVISLQRCVINLGST</INSDSeq_sequence>
    </INSDSeq>
</SequenceData>
```

Location Formats

Location operators for nucleotide sequences

- Three location operators are available for use in DNA and RNA sequences: “join,” “order,” and “complement”
- Use of the join location operators implies that the nucleotide residues described by the location descriptors are physically brought into contact by biological processes (ST.26, paragraph 68)
- Locations using “join” and “order” must have at least two comma-separated location descriptors
- “complement” can be used in combination with “join” or “order

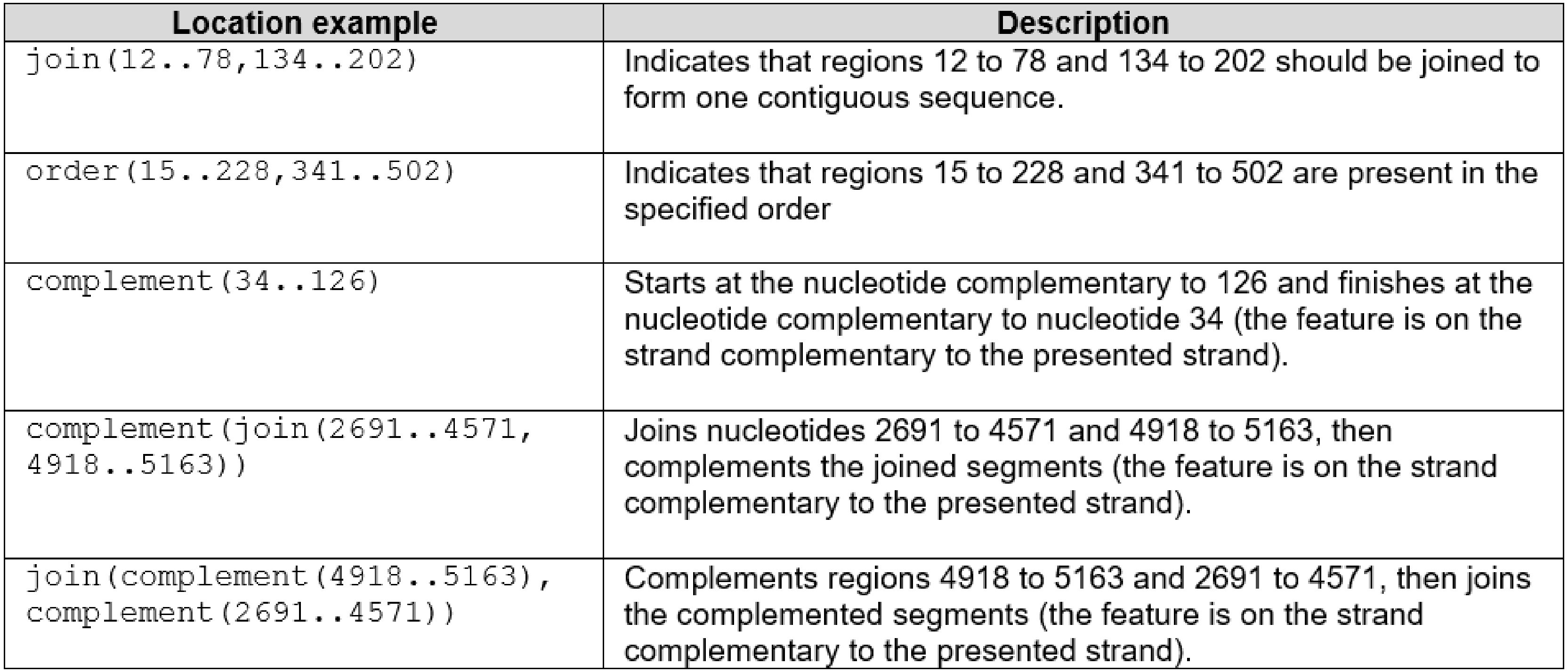


Location syntax	Location description
`join(location,location,...,location)`	The indicated locations are joined (placed end-to-end) to form one contiguous sequence.
`order(location,location,...,location)`	The elements are found in the specified order but nothing is implied about whether joining those elements is reasonable.
`complement(location)`	Indicates that the feature is located on the strand complementary to the sequence span specified by the location descriptor, when read in the 5’ to 3’ direction or in the direction that mimics the 5’ to 3’ direction.

WIPO Standard ST.26, paragraph 67

Location Formats

Location operators for nucleotide sequences



Location example	Description
`join(12..78,134..202)`	Indicates that regions 12 to 78 and 134 to 202 should be joined to form one contiguous sequence.
`order(15..228,341..502)`	Indicates that regions 15 to 228 and 341 to 502 are present in the specified order
`complement(34..126)`	Starts at the nucleotide complementary to 126 and finishes at the nucleotide complementary to nucleotide 34 (the feature is on the strand complementary to the presented strand).
`complement(join(2691..4571, 4918..5163))`	Joins nucleotides 2691 to 4571 and 4918 to 5163, then complements the joined segments (the feature is on the strand complementary to the presented strand).
`join(complement(4918..5163), complement(2691..4571))`	Complements regions 4918 to 5163 and 2691 to 4571, then joins the complemented segments (the feature is on the strand complementary to the presented strand).

WIPO Standard ST.26, paragraph 70(b)

Questions?

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

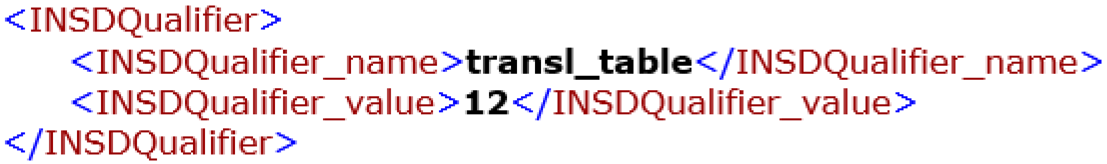
Qualifier Value Formats

Qualifier Values

- Qualifiers further define features

Qualifier Values

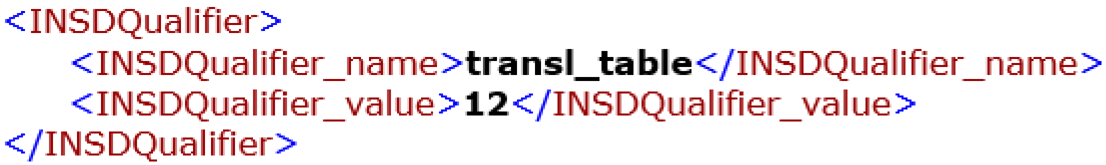
- Qualifiers further define features
- Qualifiers consist of a qualifier name and often a qualifier value



```
<INSDQualifier>
    <INSDQualifier_name>transl_table</INSDQualifier_name>
    <INSDQualifier_value>12</INSDQualifier_value>
</INSDQualifier>
```

Qualifier Values

- Qualifiers further define features
- Qualifiers consist of a qualifier name and often a qualifier value



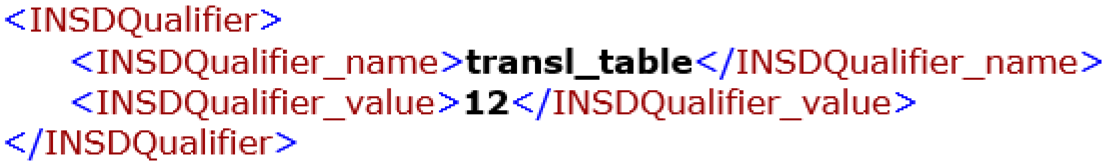
```
<INSDQualifier>
    <INSDQualifier_name>transl_table</INSDQualifier_name>
    <INSDQualifier_value>12</INSDQualifier_value>
</INSDQualifier>
```

- Each feature key will have a list of qualifiers that are permitted for that feature. Some feature keys have mandatory qualifiers.

Qualifier Values

- Qualifiers further define features

- Qualifiers consist of a qualifier name and often a qualifier value



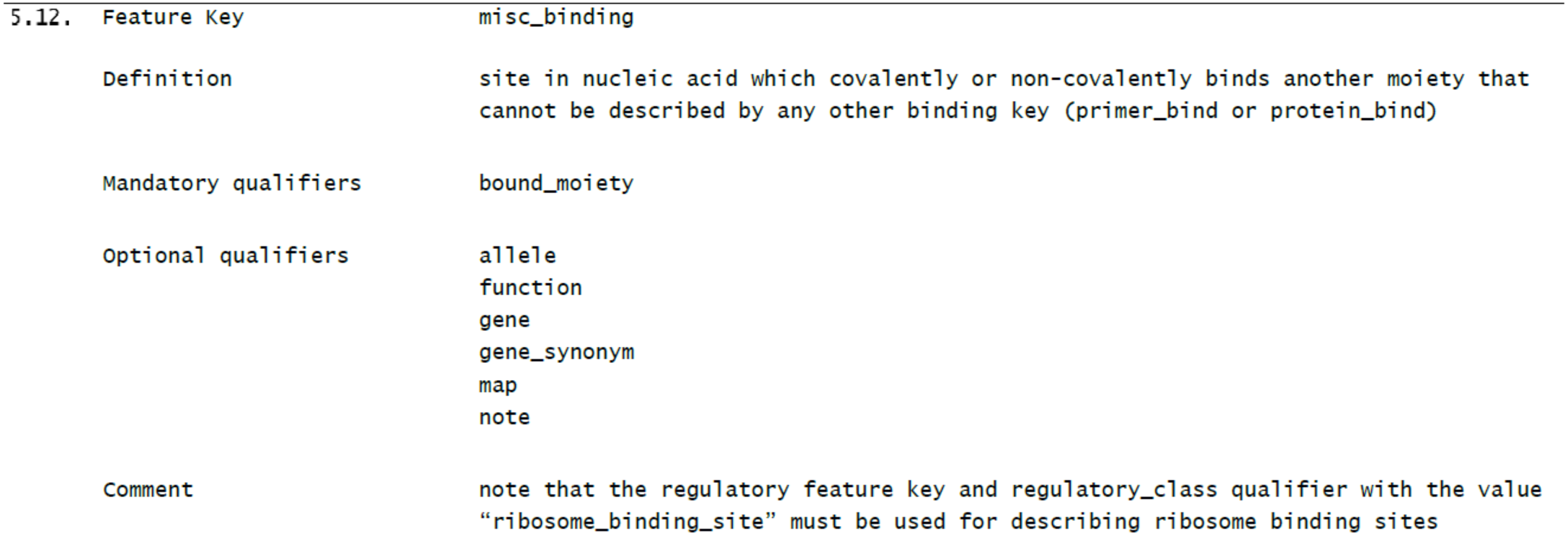
```
<INSDQualifier>
    <INSDQualifier_name>transl_table</INSDQualifier_name>
    <INSDQualifier_value>12</INSDQualifier_value>
</INSDQualifier>
```

- Each feature key will have a list of qualifiers that are permitted for that feature. Some feature keys have mandatory qualifiers.

- Feature keys and their permitted qualifiers are listed in ST.26, Annex I, Section 5 (nucleotide sequences) and Section 7 (amino acid sequences).

Qualifier Values

- Example - the Feature Key “misc_binding” has one mandatory qualifier, “bound_moiety,” and 6 optional qualifiers:



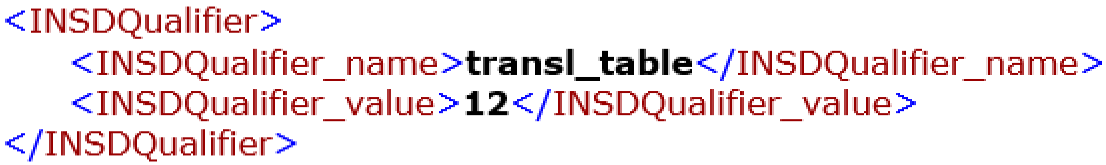
5.12.	Feature Key	misc_binding
	Definition	site in nucleic acid which covalently or non-covalently binds another moiety that cannot be described by any other binding key (primer_bind or protein_bind)
	Mandatory qualifiers	bound_moiety
	Optional qualifiers	allele function gene gene_synonym map note
	Comment	note that the regulatory feature key and regulatory_class qualifier with the value “ribosome_binding_site” must be used for describing ribosome binding sites

(ST.26, Annex I, Section 5.12)

Qualifier Values

- Qualifiers further define features

- Qualifiers consist of a qualifier name and often a qualifier value



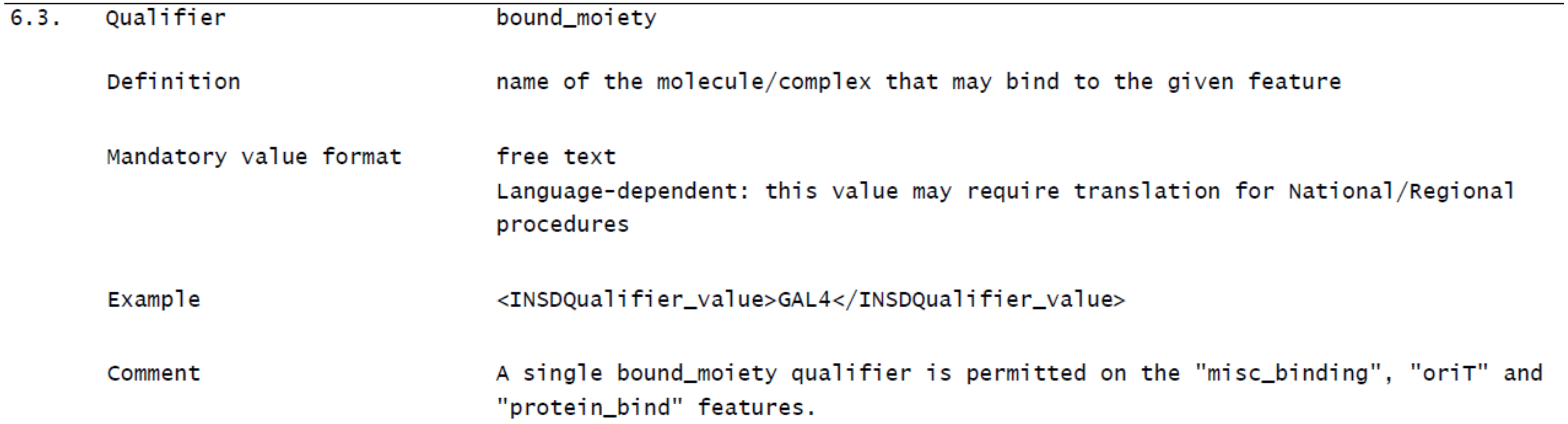
```
<INSDQualifier>
    <INSDQualifier_name>transl_table</INSDQualifier_name>
    <INSDQualifier_value>12</INSDQualifier_value>
</INSDQualifier>
```

- Each feature key will have a list of qualifiers that are permitted for that feature. Some feature keys have mandatory qualifiers.

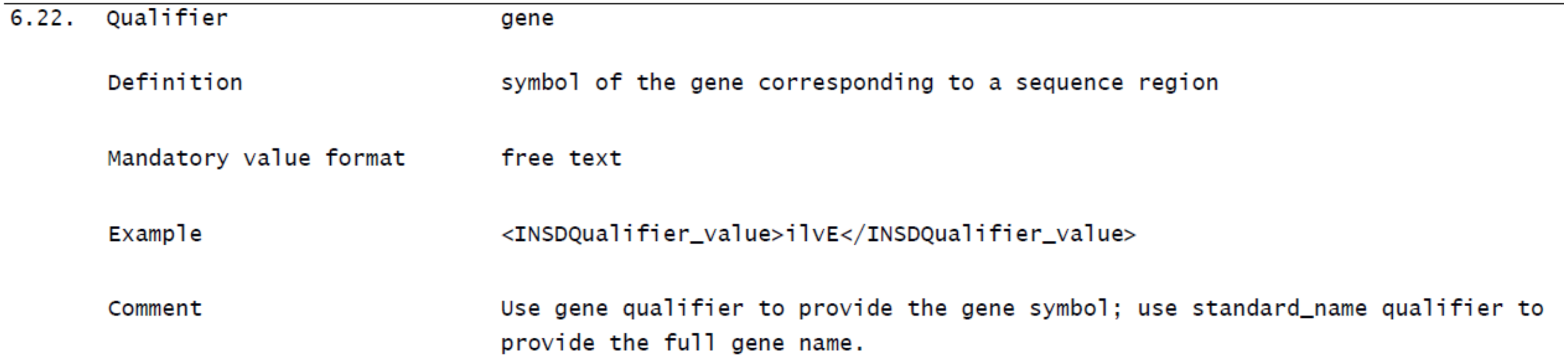
- Feature keys and their permitted qualifiers are listed in ST.26, Annex I, Section 5 (nucleotide sequences) and Section 7 (amino acid sequences).

- Qualifiers for nucleotide sequences along with their descriptions are listed in ST.26, Annex I, Section 6 (nucleotide sequences) and Section 8 (amino acid sequences).

Qualifier Values



6.3.	Qualifier	bound_moiety
	Definition	name of the molecule/complex that may bind to the given feature
	Mandatory value format	free text Language-dependent: this value may require translation for National/Regional procedures
	Example	<INSDQualifier_value>GAL4</INSDQualifier_value>
	Comment	A single bound_moiety qualifier is permitted on the "misc_binding", "oriT" and "protein_bind" features.



6.22.	Qualifier	gene
	Definition	symbol of the gene corresponding to a sequence region
	Mandatory value format	free text
	Example	<INSDQualifier_value>ilvE</INSDQualifier_value>
	Comment	Use gene qualifier to provide the gene symbol; use standard_name qualifier to provide the full gene name.

(ST.26, Annex I, Section 6.3 and 6.22)

Qualifier Values

Format types

- Qualifier values have several format types:

  1. Qualifiers with predefined value choices;
  2. Qualifiers with a defined value format;
  3. Qualifiers where the value is a sequence;
  4. Qualifiers with NO value;
  5. Qualifiers with “free text” values

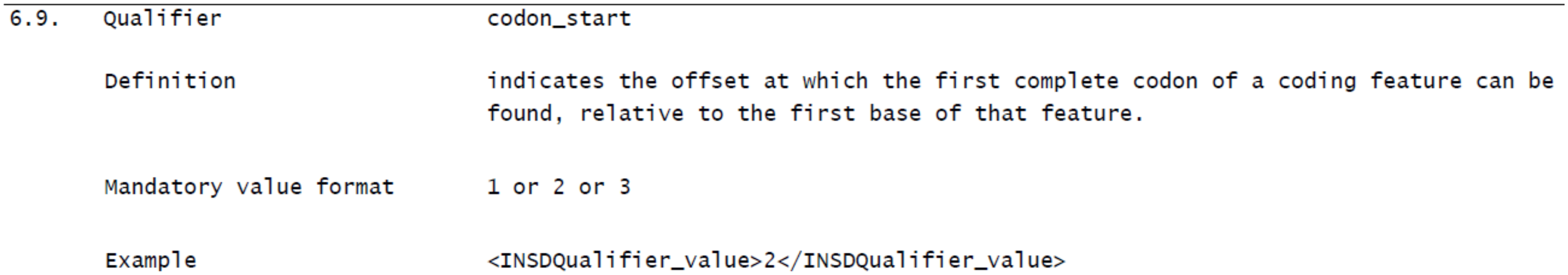
     - a subset of “free text” qualifier values are categorized as “language dependent”

Qualifier Values

Format types – Predefined value choices

- Qualifiers with predefined value choices
- Examples:

“codon_start” – values can be “1”, “2”, or “3”

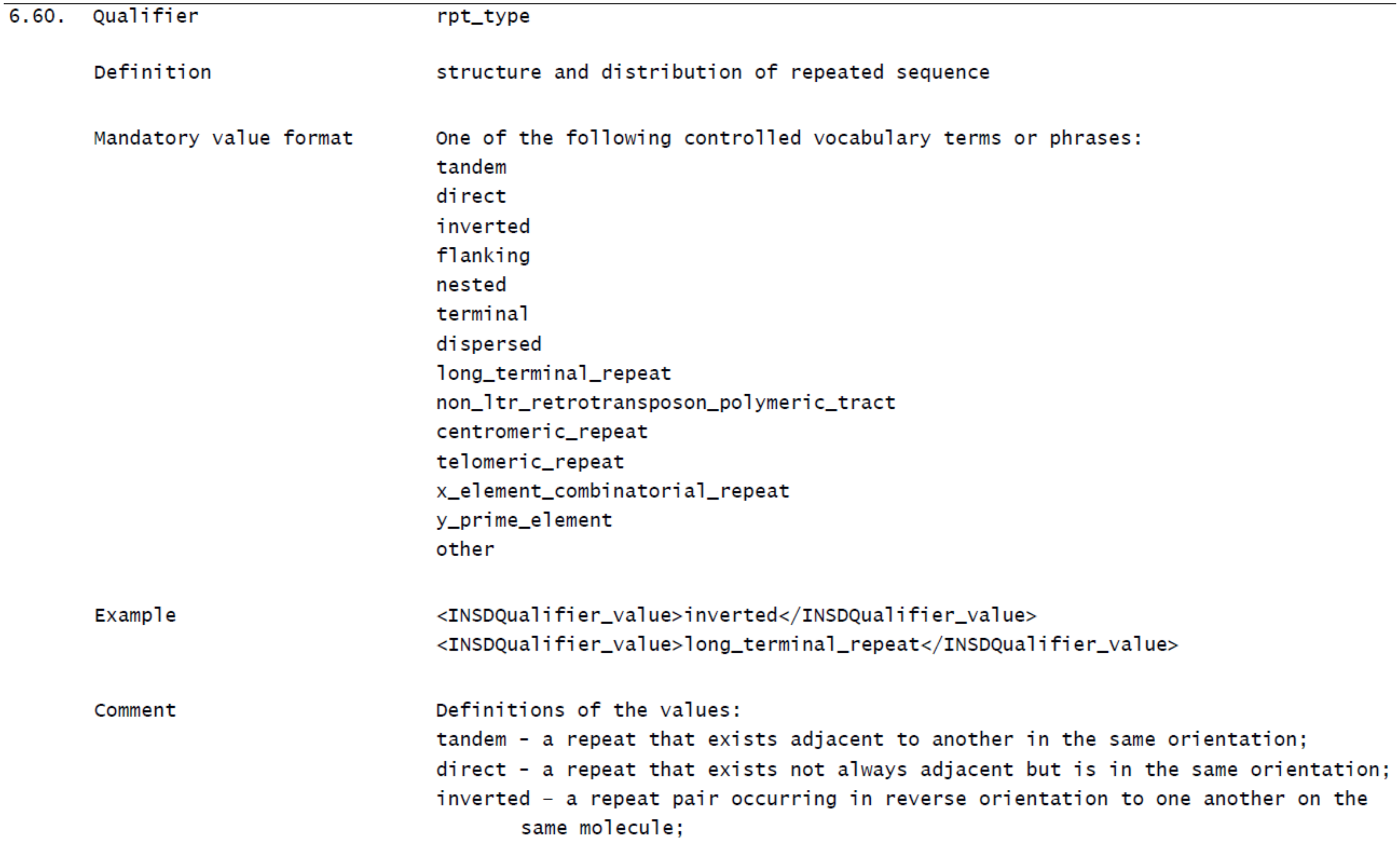


6.9.	Qualifier	codon_start
	Definition	indicates the offset at which the first complete codon of a coding feature can be found, relative to the first base of that feature.
	Mandatory value format	1 or 2 or 3
	Example	<INSDQualifier_value>2</INSDQualifier_value>

Qualifier Values

Format types – Predefined value choices

- Examples: “rpt_type” has a limited set of value choices

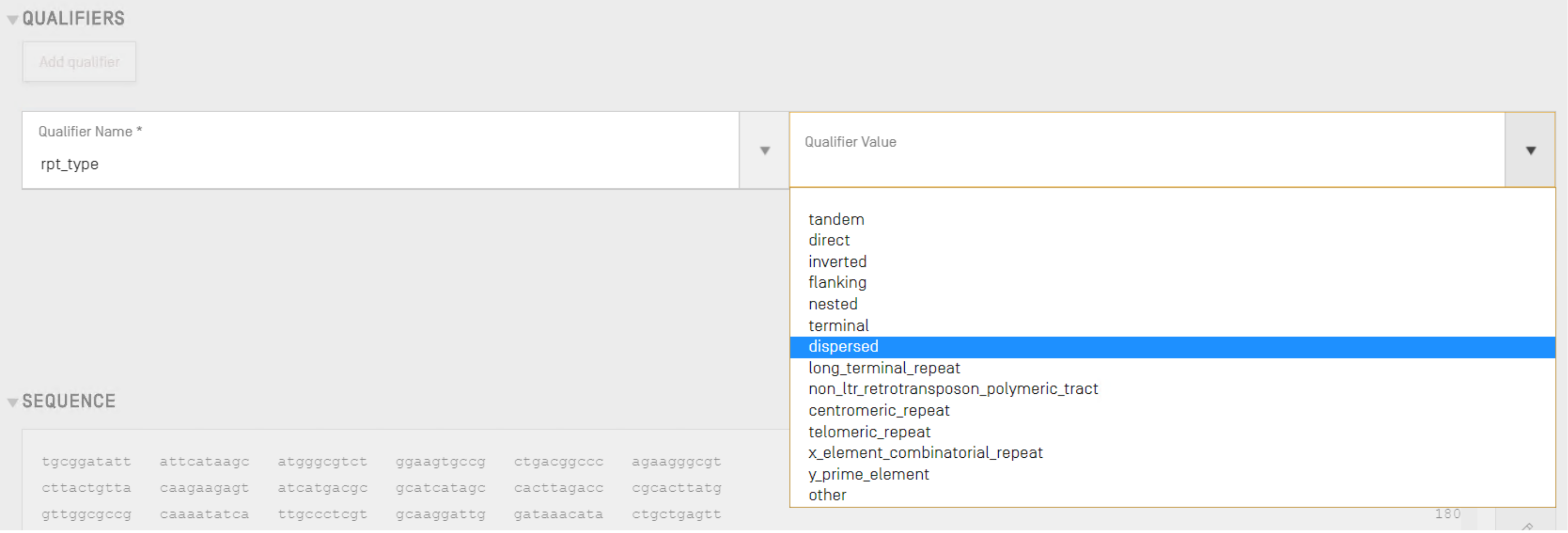


6.60.	Qualifier	rpt_type
	Definition	structure and distribution of repeated sequence
	Mandatory value format	One of the following controlled vocabulary terms or phrases: tandem direct inverted flanking nested terminal dispersed long_terminal_repeat non_ltr_retrotransposon_polymeric_tract centromeric_repeat telomeric_repeat x_element_combinatorial_repeat y_prime_element other
	Example	<INSDQualifier_value>inverted</INSDQualifier_value> <INSDQualifier_value>long_terminal_repeat</INSDQualifier_value>
	Comment	Definitions of the values: tandem - a repeat that exists adjacent to another in the same orientation; direct - a repeat that exists not always adjacent but is in the same orientation; inverted – a repeat pair occurring in reverse orientation to one another on the same molecule;

Qualifier Values

Format types – Predefined value choices

- For qualifiers with predefined value choices, WIPO Sequence will present all permitted values in a prepopulated drop-down list:



Qualifier Name *
rpt_type

Qualifier Value

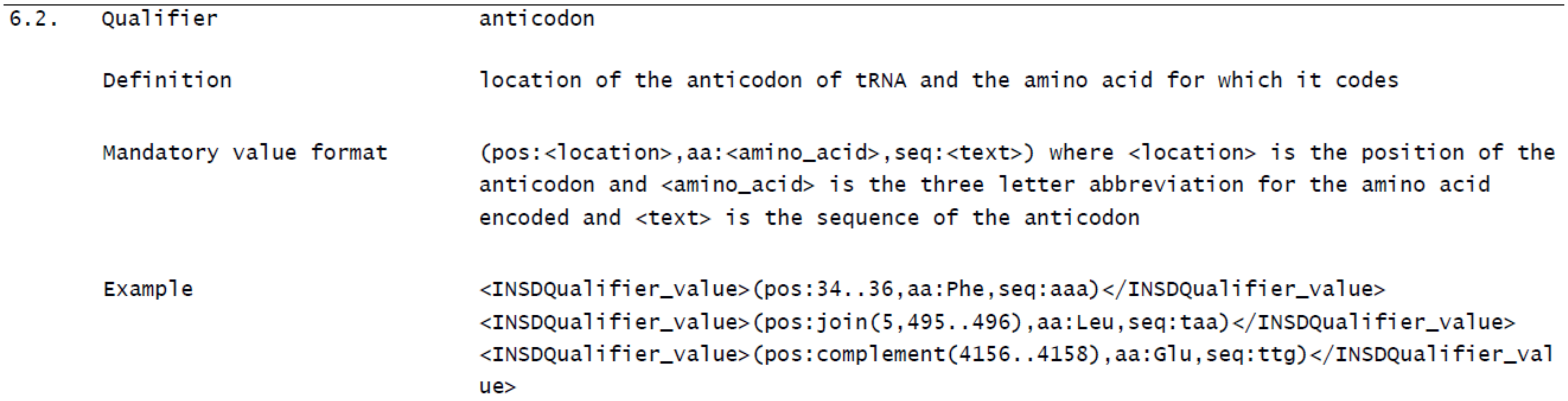
tandem
direct
inverted
flanking
nested
terminal
dispersed
long_terminal_repeat
non_ltr_retrotransposon_polymeric_tract
centromeric_repeat
telomeric_repeat
x_element_combinatorial_repeat
y_prime_element
other

```
tgcggatatt attcataagc atgggcgtct ggaagtgccg ctgacggccc agaagggcgt
cttactgtta caagaagagt atcatgacgc gcatcatagc cacttagacc cgcacttatg
gttggcgccg caaaatatca ttgccctcgt gcaaggattg gataaacata ctgctgagtt 180
```

Qualifier Values

Format types – Defined value format

- Qualifiers with a defined value format
- Example: “anticodon” – value must be in the format “(pos:<location>,aa:<amino_acid>,seq:<text>)”

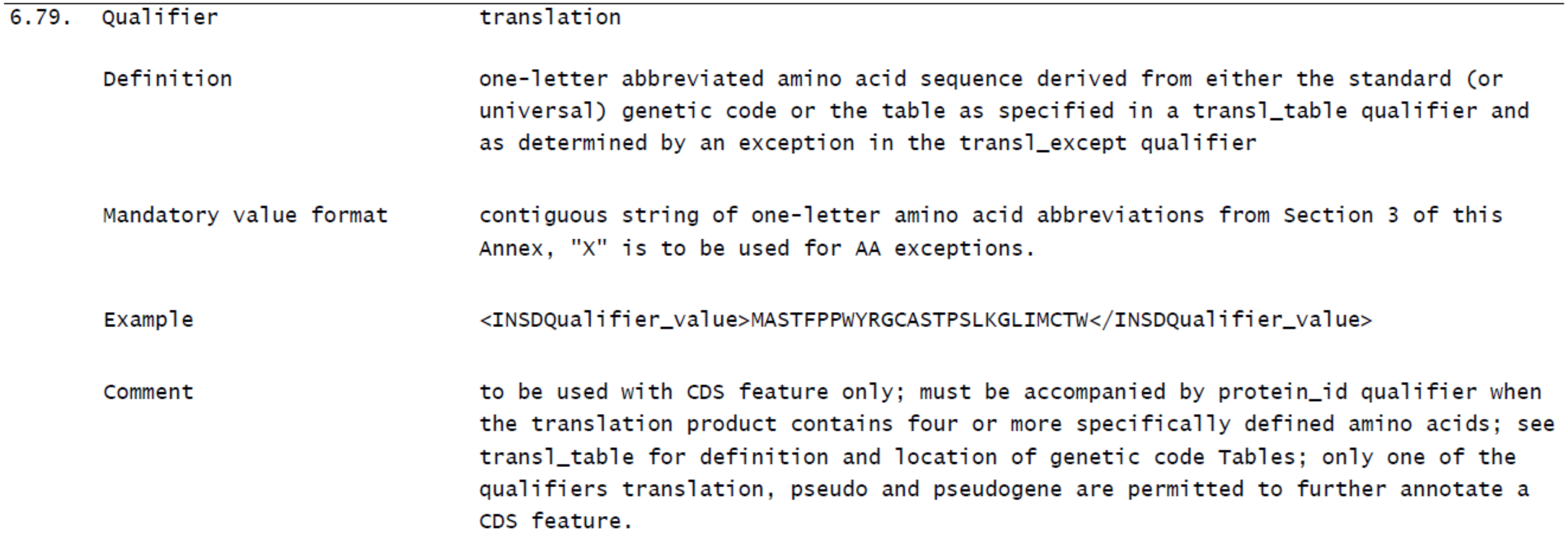


6.2.	Qualifier	anticodon
	Definition	location of the anticodon of tRNA and the amino acid for which it codes
	Mandatory value format	(pos:<location>,aa:<amino_acid>,seq:<text>) where <location> is the position of the anticodon and <amino_acid> is the three letter abbreviation for the amino acid encoded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

Qualifier Values

Format types - Sequences

- Qualifiers where the value is a sequence
- Example: “translation” – value must be a sequence using one letter amino acid abbreviations

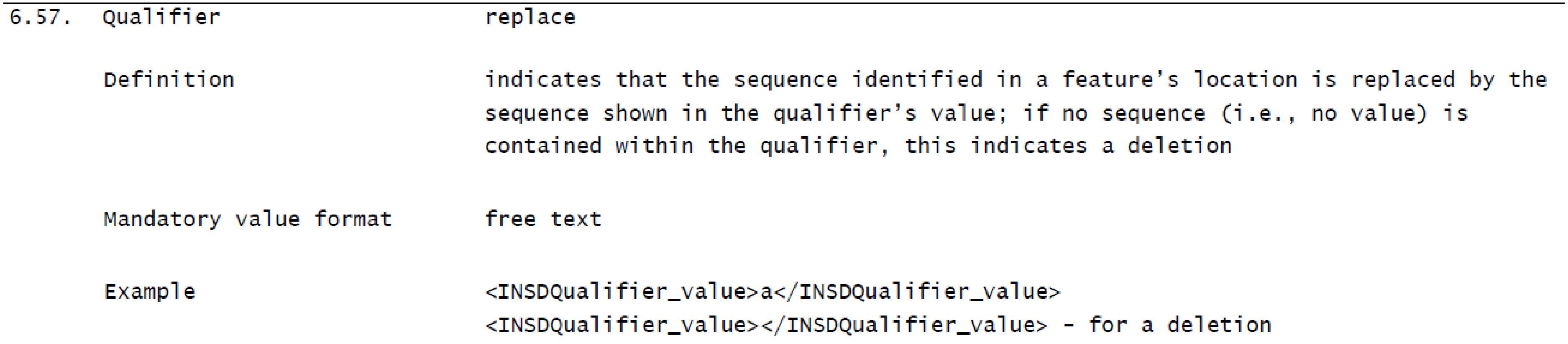


6.79.	Qualifier	translation
	Definition	one-letter abbreviated amino acid sequence derived from either the standard (or universal) genetic code or the table as specified in a transl_table qualifier and as determined by an exception in the transl_except qualifier
	Mandatory value format	contiguous string of one-letter amino acid abbreviations from Section 3 of this Annex, "X" is to be used for AA exceptions.
	Example	<INSDQualifier_value>MASTFPPWYRGCASTPSLKGLIMCTW</INSDQualifier_value>
	Comment	to be used with CDS feature only; must be accompanied by protein_id qualifier when the translation product contains four or more specifically defined amino acids; see transl_table for definition and location of genetic code Tables; only one of the qualifiers translation, pseudo and pseudogene are permitted to further annotate a CDS feature.

Qualifier Values

Format types - Sequences

- Qualifier “replace” – value can be a single nucleotide residue, a sequence of residues, or empty
- An empty value for “replace” signifies a deletion of the residue indicated in the corresponding feature



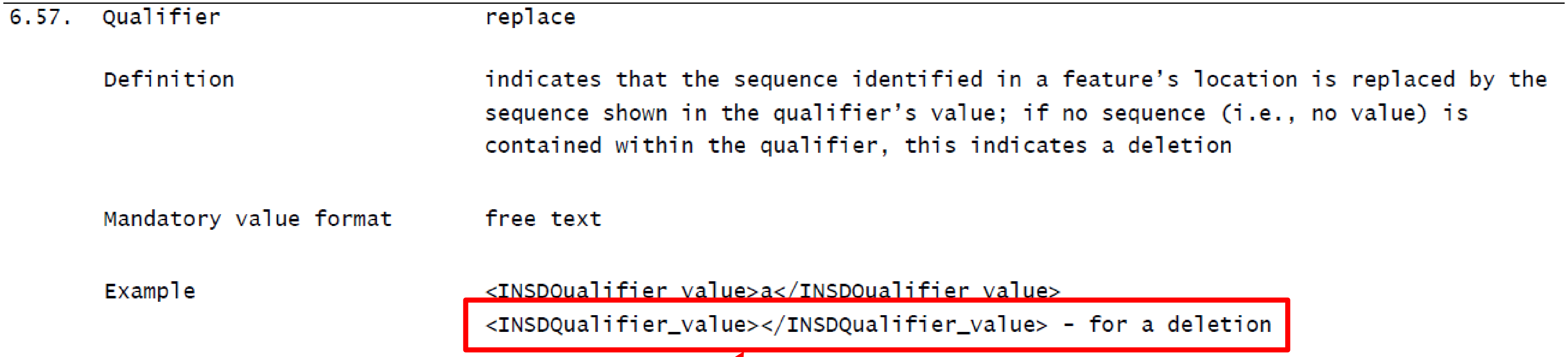
6.57.	Qualifier	replace
	Definition	indicates that the sequence identified in a feature’s location is replaced by the sequence shown in the qualifier’s value; if no sequence (i.e., no value) is contained within the qualifier, this indicates a deletion
	Mandatory value format	free text
	Example	<INSDQualifier_value>a</INSDQualifier_value> <INSDQualifier_value></INSDQualifier_value> - for a deletion

Qualifier Values

Format types - Sequences

- Qualifier “replace” – value can be a single nucleotide residue, a sequence of residues, or empty

- An empty value for “replace” signifies a deletion of the residue indicated in the corresponding feature



6.57.	Qualifier	replace
	Definition	indicates that the sequence identified in a feature’s location is replaced by the sequence shown in the qualifier’s value; if no sequence (i.e., no value) is contained within the qualifier, this indicates a deletion
	Mandatory value format	free text
	Example	<INSDQualifier_value>a</INSDQualifier_value> <INSDQualifier_value></INSDQualifier_value> - for a deletion

“Empty value”

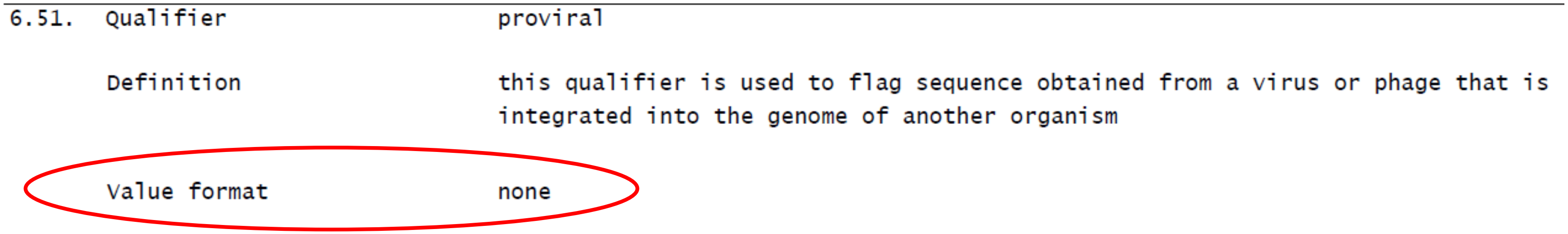
Qualifier Values

Format types - NO Value

- Qualifiers with NO value

- Examples:

“environmental_sample” “germline” “macronuclear” “proviral”



6.51.	Qualifier	proviral
	Definition	this qualifier is used to flag sequence obtained from a virus or phage that is integrated into the genome of another organism
	Value format	none

- WIPO Sequence will not allow addition of a value for qualifiers with no value

- These qualifiers must not have an empty “INSDQualifier_value” element

Qualifier Values

Format types – "Free Text"

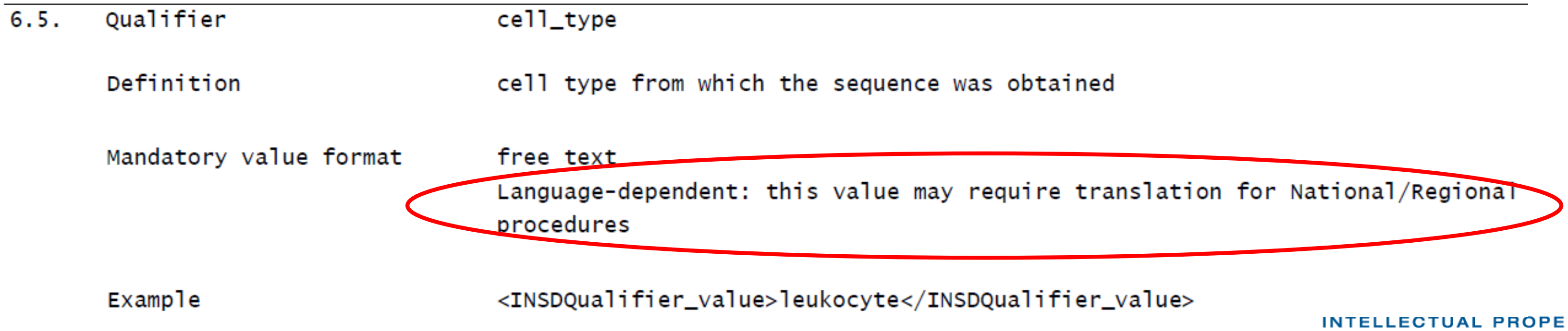
- Many qualifiers have a "free text" value format
- ST.26, paragraph 3(n) defines "free text" as "…*a type of value format for certain qualifiers, presented in the form of a descriptive text phrase or other specified format (as indicated in Annex I).*"
- Free text qualifier values are limited to 1000 characters (ST.26, pgh. 86)
- A subset of qualifiers with "free text" value formats are "*language-dependent*"
- "Language-dependent" free text "may require translation for national, regional, or international procedures."   ST.26, paragraph 3(o)

Qualifier Values

Format types – “Free Text”

- Qualifiers that have “language-dependent free text” values may require translation

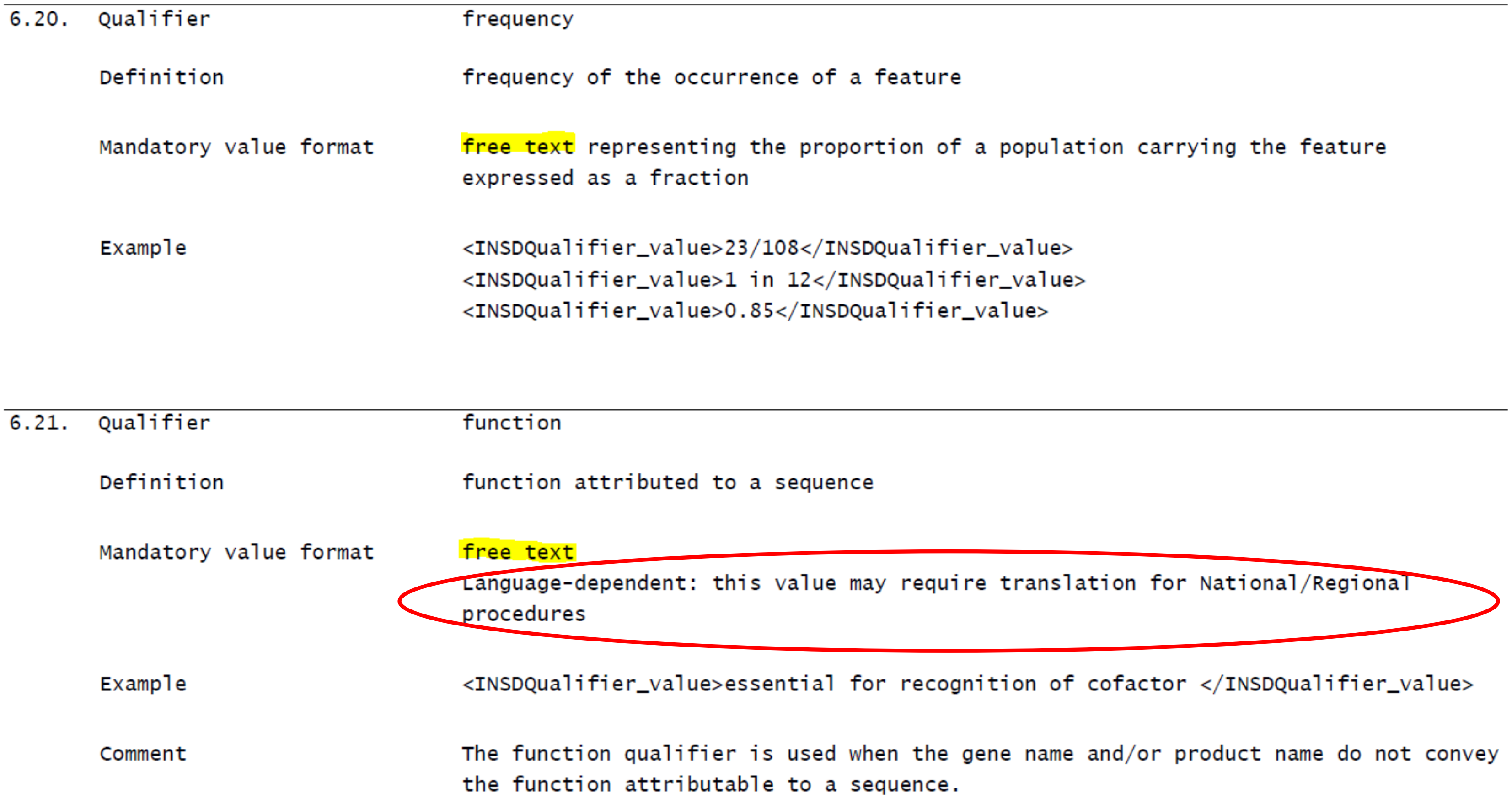
- How do you determine if a qualifier with a free text value format is language-dependent?
  - ST.26, Annex I, Section 6, Table 5 lists all nucleotide sequence qualifiers with a language-dependent free text value format
  - ST.26, Annex I, Section 8, Table 6 lists all amino acid sequence qualifiers with a language-dependent free text value format
  - Look at the “Mandatory value format” in the qualifier description



6.5.	Qualifier	cell_type
	Definition	cell type from which the sequence was obtained
	Mandatory value format	free text Language-dependent: this value may require translation for National/Regional procedures
	Example	<INSDQualifier_value>leukocyte</INSDQualifier_value>

Qualifier Values

Format types – “Free Text”



6.20.	Qualifier	frequency
	Definition	frequency of the occurrence of a feature
	Mandatory value format	free text representing the proportion of a population carrying the feature expressed as a fraction
	Example	<INSDQualifier_value>23/108</INSDQualifier_value> <INSDQualifier_value>1 in 12</INSDQualifier_value> <INSDQualifier_value>0.85</INSDQualifier_value>
6.21.	Qualifier	function
	Definition	function attributed to a sequence
	Mandatory value format	free text Language-dependent: this value may require translation for National/Regional procedures
	Example	<INSDQualifier_value>essential for recognition of cofactor </INSDQualifier_value>
	Comment	The function qualifier is used when the gene name and/or product name do not convey the function attributable to a sequence.

Qualifier Values

Format types – “Language-Dependent Free Text”

- “Language-dependent free text” qualifier values may be provided in two languages in the sequence listing XML: English and one other non-English language (ST.26, paragraph 87)

Qualifier Values

Format types – “Language-Dependent Free Text”

- “Language-dependent free text” qualifier values may be provided in two languages in the sequence listing XML: English and one other non-English language (ST.26, paragraph 87)

- A language-dependent qualifier value in English must be provided in an INSDQualifier_value element

Qualifier Values

Format types – “Language-Dependent Free Text”

- “Language-dependent free text” qualifier values may be provided in two languages in the sequence listing XML: English and one other non-English language (ST.26, paragraph 87)

- A language-dependent qualifier value in English must be provided in an INSDQualifier_value element

- A language-dependent qualifier value in any language other than English must be provided in a NonEnglishQualifier_value element

Qualifier Values

Format types – “Language-Dependent Free Text”

- “Language-dependent free text” qualifier values may be provided in two languages in the sequence listing XML: English and one other non-English language (ST.26, paragraph 87)

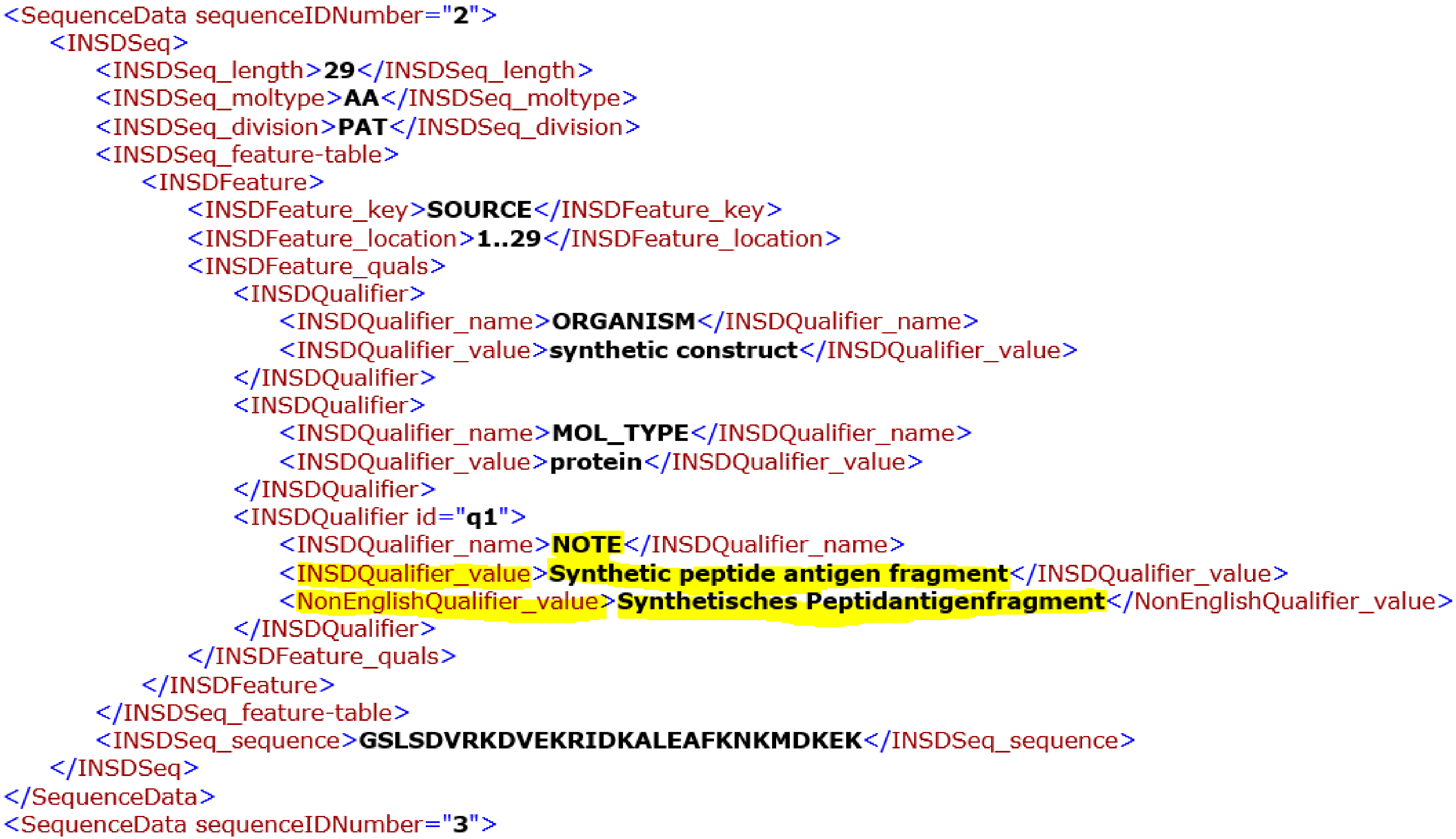
- A language-dependent qualifier value in English must be provided in an INSDQualifier_value element

- A language-dependent qualifier value in any language other than English must be provided in a NonEnglishQualifier_value element

- A NonEnglishQualifier_value element is permitted ONLY for qualifiers with language-dependent free text value format (ST.26, paragraph 87(b))

Qualifier Values

Format types – “Language-Dependent Free Text”



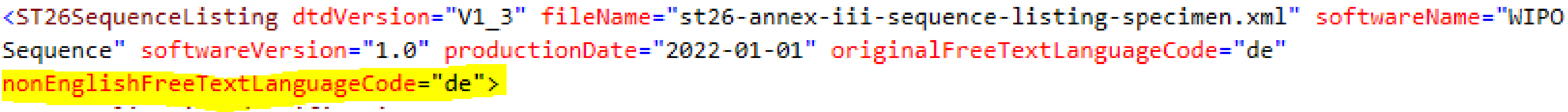
```
<SequenceData sequenceIDNumber="2">
    <INSDSeq>
        <INSDSeq_length>29</INSDSeq_length>
        <INSDSeq_moltype>AA</INSDSeq_moltype>
        <INSDSeq_division>PAT</INSDSeq_division>
        <INSDSeq_feature-table>
            <INSDFeature>
                <INSDFeature_key>SOURCE</INSDFeature_key>
                <INSDFeature_location>1..29</INSDFeature_location>
                <INSDFeature_quals>
                    <INSDQualifier>
                        <INSDQualifier_name>ORGANISM</INSDQualifier_name>
                        <INSDQualifier_value>synthetic construct</INSDQualifier_value>
                    </INSDQualifier>
                    <INSDQualifier>
                        <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
                        <INSDQualifier_value>protein</INSDQualifier_value>
                    </INSDQualifier>
                    <INSDQualifier id="q1">
                        <INSDQualifier_name>NOTE</INSDQualifier_name>
                        <INSDQualifier_value>Synthetic peptide antigen fragment</INSDQualifier_value>
                        <NonEnglishQualifier_value>Synthetisches Peptidantigenfragment</NonEnglishQualifier_value>
                    </INSDQualifier>
                </INSDFeature_quals>
            </INSDFeature>
        </INSDSeq_feature-table>
        <INSDSeq_sequence>GSLSDVRKDVEKRIDKALEAFKNKMDKEK</INSDSeq_sequence>
    </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="3">
```

Qualifier Values

Format types – “Language-Dependent Free Text”

- When a sequence listing XML contains non-English qualifier values:

  1. the root element of the sequence listing XML must contain a “nonEnglishFreeTextLanguageCode” attribute with an appropriate 2-letter language code abbreviation (ST.26 paragraphs 43 and 87(b));



```
<ST26SequenceListing dtdVersion="V1_3" fileName="st26-annex-iii-sequence-listing-specimen.xml" softwareName="WIPO
Sequence" softwareVersion="1.0" productionDate="2022-01-01" originalFreeTextLanguageCode="de"
nonEnglishFreeTextLanguageCode="de">
```

Qualifier Values

Format types – “Language-Dependent Free Text”

- When a sequence listing XML contains non-English qualifier values:

  1. the root element of the sequence listing XML must contain a “nonEnglishFreeTextLanguageCode attribute with an appropriate 2-letter language code abbreviation (ST.26 paragraphs 43 and 87(b));

  2. ALL language-dependent qualifiers in the sequence listing must have values in the language indicated in the “nonEnglishFreeTextLanguageCode” attribute;

Qualifier Values

Format types – “Language-Dependent Free Text”

- When a sequence listing XML contains non-English qualifier values:

1. the root element of the sequence listing XML must contain a “nonEnglishFreeTextLanguageCode attribute with an appropriate 2-letter language code abbreviation (ST.26 paragraphs 43 and 87(b));

2. ALL language-dependent qualifiers in the sequence listing must have values in the language indicated in the “nonEnglishFreeTextLanguageCode” attribute;

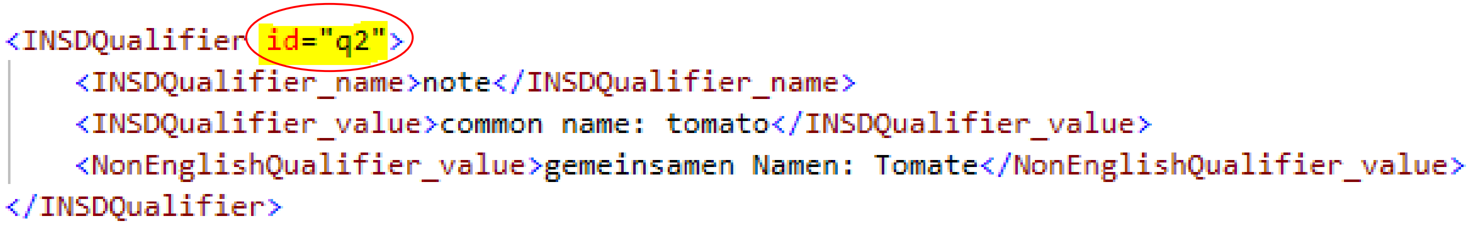
3. Where NonEnglishQualifier_value and INSDQualifier_value are both present for a single qualifier, the information contained in the two elements must be equivalent (ST.26 paragraph 87(c)).

Qualifier Values

Format types – “Language-Dependent Free Text”

- INSDQualifier “id” attribute – what is it?

- ST.26 paragraph 87(d) states, *“For language-dependent qualifiers, the INSDQualifier element may include an optional attribute id. The value of this attribute must be in the format "q" followed by a positive integer, e.g. "q23", and must be unique to one INSDQualifier element, i.e. the attribute value must only be used once in a sequence listing file.*”



```
<INSDQualifier id="q2">
    <INSDQualifier_name>note</INSDQualifier_name>
    <INSDQualifier_value>common name: tomato</INSDQualifier_value>
    <NonEnglishQualifier_value>gemeinsamen Namen: Tomate</NonEnglishQualifier_value>
</INSDQualifier>
```

Qualifier Values

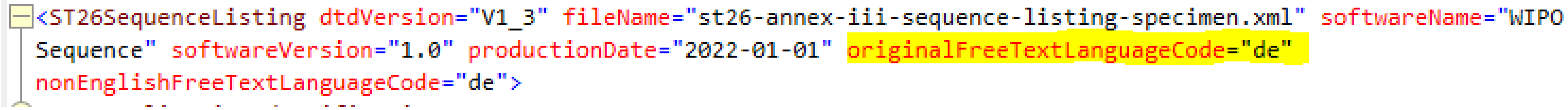
Format types – “Language-Dependent Free Text”

- INSDQualifier “id” attributes:
  - Uniquely identify qualifier values that may require translation for export to an XLIFF file by WIPO Sequence;
  - Optional;
  - Only permitted for language-dependent qualifiers;
  - Must be unique in a sequence listing;
  - Automatically added in an XML generated by WIPO Sequence.

Qualifier Values

Format types – “Language-Dependent Free Text”

- The “originalFreeTextLanguageCode” attribute



```
<ST26SequenceListing dtdVersion="V1_3" fileName="st26-annex-iii-sequence-listing-specimen.xml" softwareName="WIPO
Sequence" softwareVersion="1.0" productionDate="2022-01-01" originalFreeTextLanguageCode="de"
nonEnglishFreeTextLanguageCode="de">
```

- ST.26 paragraph 43 defines the “originalFreeTextLanguageCode” attribute as, “The language code…for the single original language in which the language-dependent free text qualifiers were prepared.”

- This attribute is OPTIONAL

Special Situations:
DNA/RNA hybrid molecules
Uracil in DNA and Thymine in RNA

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

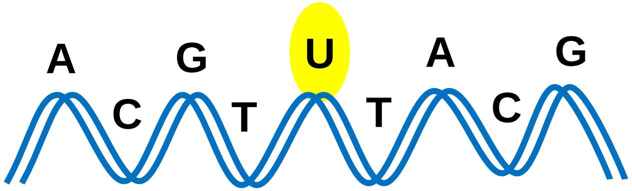
Special Situations

- Reminder: the “u” symbol for uracil is not permitted in ST.26
- In DNA, “t” is thymine
- In RNA, “t” is uracil
- Two scenarios to consider:
  1. DNA molecule with uracil nucleobase or RNA molecule with thymine nucleobase;
  2. DNA/RNA hybrid molecule

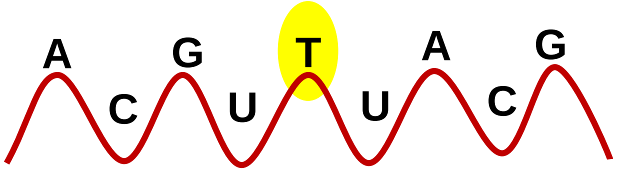
Special Situations

Uracil in DNA and Thymine in RNA

- If a sequence has a DNA backbone with a uracil nucleobase, OR an RNA backbone with a thymine nucleobase – describe as a “modified nucleotide”



DNA



RNA

<u>ST.26 Paragraph 14 applies</u>:

“14. The symbol “t” will be construed as thymine in DNA and uracil in RNA. Uracil in DNA or thymine in RNA is considered a <u>modified nucleotide</u> and must be further described in the feature table as provided by paragraph 19.”

Special Situations

DNA and RNA Sequences

An application discloses the following RNA sequence:

5′-cgucccacgugtccgaggua-3′

- Note the “thymine” at position 12. This residue must be annotated as a modified nucleotide.

- ST.26 paragraph 19 states: *Uracil in DNA or thymine in RNA are considered modified nucleotides and must be represented in the sequence as “t” and be further described in the feature table using the feature key “modified_base”, the qualifier “mod_base” with “OTHER” as the qualifier value and the qualifier “note” with “uracil” or “thymine”, respectively, as the qualifier value.*

Special Situations

DNA and RNA Sequences

An application discloses the following RNA sequence:

5′-cgucccacgugtccgaggua-3′

- Note the “thymine” at position 12. This residue must be annotated as a modified nucleotide.

- ST.26 paragraph 19 states: *Uracil in DNA or thymine in RNA are considered modified nucleotides and must be represented in the sequence as “t” and be further described in the feature table using the feature key “modified_base”, the qualifier “mod_base” with “OTHER” as the qualifier value and the qualifier “note” with “uracil” or “thymine”, respectively, as the qualifier value.*

Special Situations

DNA and RNA Sequences

An application discloses the following RNA sequence:

5′-cgucccacgugtccgaggua-3′

✓ All uracil residues must be represented by the symbol “t”. Therefore, the sequence must be represented in the sequence listing as:

cgtcccacgtgtccgaggta

Special Situations

DNA and RNA Sequences

An application discloses the following RNA sequence:

`5′-cgucccacgugtccgaggua-3′`

✓ All uracil residues must be represented by the symbol “t”. Therefore, the sequence must be represented in the sequence listing as:

`cgtcccacgtgtccgaggta`

✓ Feature key “modified_base” with location “12”

Special Situations

DNA and RNA Sequences

An application discloses the following RNA sequence:

5′-cgucccacgugtccgaggua-3′

✓ All uracil residues must be represented by the symbol “t”. Therefore, the sequence must be represented in the sequence listing as:

cgtcccacgtgtccgaggta

✓ Feature key “modified_base” with location “12”

✓ Qualifier “mod_base” with value “OTHER”

Special Situations

DNA and RNA Sequences

An application discloses the following RNA sequence:

```
5′-cgucccacgugtccgaggua-3′
```

✓ All uracil residues must be represented by the symbol “t”. Therefore, the sequence must be represented in the sequence listing as:

```
cgtcccacgtgtccgaggta
```

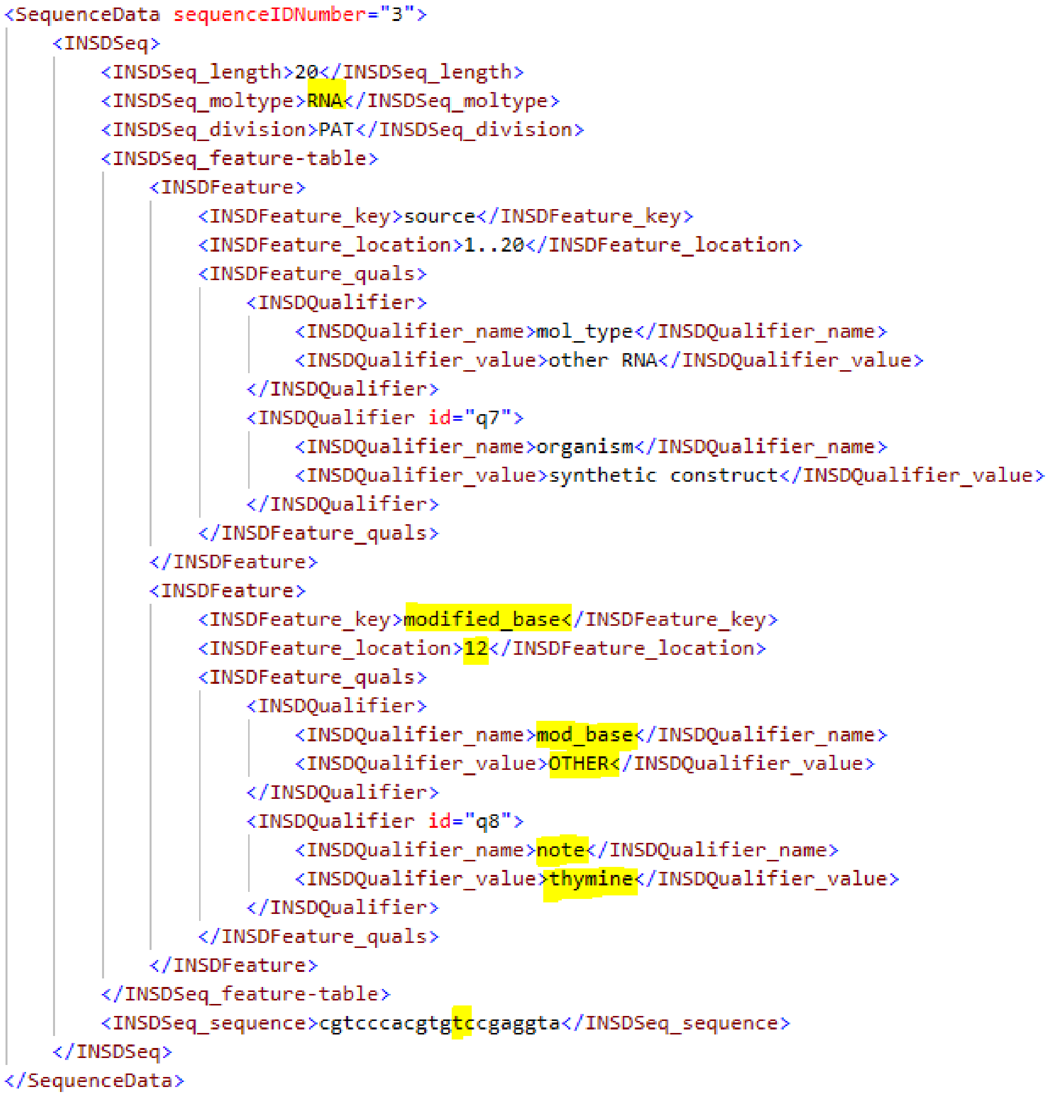
✓ Feature key “modified_base” with location “12”

✓ Qualifier “mod_base” with value “OTHER”

✓ Qualifier “note” with the value “thymine”

Special Situations

DNA and RNA Sequences

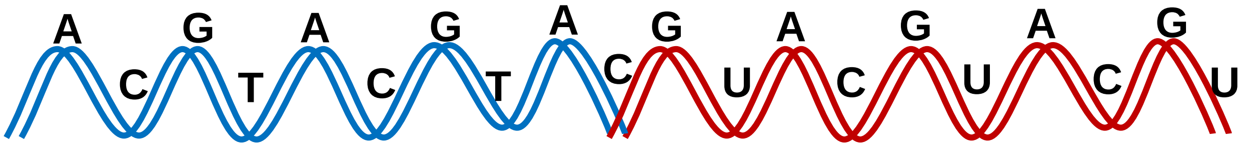


```
<SequenceData sequenceIDNumber="3">
    <INSDSeq>
        <INSDSeq_length>20</INSDSeq_length>
        <INSDSeq_moltype>RNA</INSDSeq_moltype>
        <INSDSeq_division>PAT</INSDSeq_division>
        <INSDSeq_feature-table>
            <INSDFeature>
                <INSDFeature_key>source</INSDFeature_key>
                <INSDFeature_location>1..20</INSDFeature_location>
                <INSDFeature_quals>
                    <INSDQualifier>
                        <INSDQualifier_name>mol_type</INSDQualifier_name>
                        <INSDQualifier_value>other RNA</INSDQualifier_value>
                    </INSDQualifier>
                    <INSDQualifier id="q7">
                        <INSDQualifier_name>organism</INSDQualifier_name>
                        <INSDQualifier_value>synthetic construct</INSDQualifier_value>
                    </INSDQualifier>
                </INSDFeature_quals>
            </INSDFeature>
            <INSDFeature>
                <INSDFeature_key>modified_base</INSDFeature_key>
                <INSDFeature_location>12</INSDFeature_location>
                <INSDFeature_quals>
                    <INSDQualifier>
                        <INSDQualifier_name>mod_base</INSDQualifier_name>
                        <INSDQualifier_value>OTHER</INSDQualifier_value>
                    </INSDQualifier>
                    <INSDQualifier id="q8">
                        <INSDQualifier_name>note</INSDQualifier_name>
                        <INSDQualifier_value>thymine</INSDQualifier_value>
                    </INSDQualifier>
                </INSDFeature_quals>
            </INSDFeature>
        </INSDSeq_feature-table>
        <INSDSeq_sequence>cgtcccacgtgtccgaggta</INSDSeq_sequence>
    </INSDSeq>
</SequenceData>
```

Special Situations

DNA/RNA Hybrid Sequences

- If a sequence is a hybrid molecule; i.e., part of the backbone is DNA and part of the backbone is RNA:



DNA segment

RNA segment

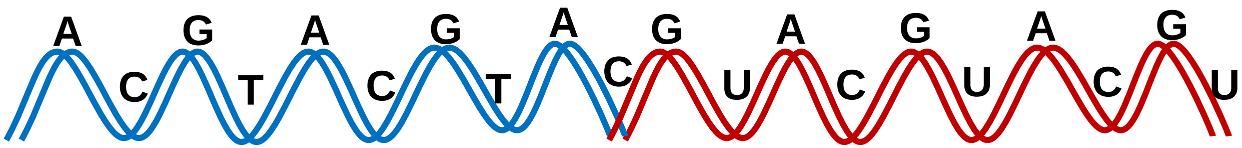
ST.26 Paragraph 55 applies:

"55. For a nucleotide sequence that contains both DNA and RNA segments of one or more nucleotides, the molecule type must be indicated as DNA. The combined DNA/RNA molecule must be further described in the feature table, using the feature key "source" and the mandatory qualifier "organism" with the value "synthetic construct" and the mandatory qualifier "mol_type" with the value "other DNA". Each DNA and RNA segment of the combined DNA/RNA molecule must be further described with the feature key "misc_feature" and the qualifier "note", which indicates whether the segment is DNA or RNA."

Special Situations

DNA/RNA Hybrid Sequences

- If a sequence is a hybrid molecule; i.e., part of the backbone is DNA and part of the backbone is RNA:



DNA segment

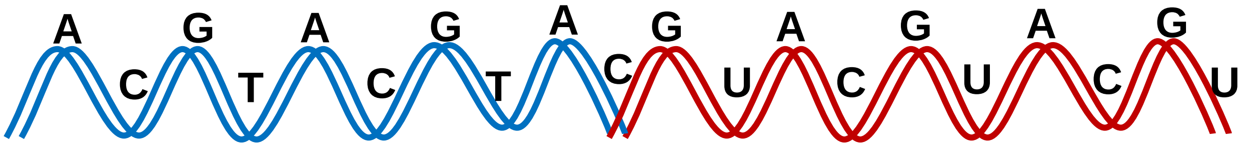
RNA segment

<u>ST.26 Paragraph 55 applies</u>:

"55. For a nucleotide sequence that contains both DNA and RNA segments of one or more nucleotides, the molecule type must be indicated as DNA. The combined DNA/RNA molecule must be further described in the feature table, using the feature key "source" and the mandatory qualifier "organism" with the value "synthetic construct" and the mandatory qualifier "mol_type" with the value "other DNA". Each DNA and RNA segment of the combined DNA/RNA molecule must be further described with the feature key "misc_feature" and the qualifier "note", which indicates whether the segment is DNA or RNA."

Special Situations
DNA/RNA Hybrid Sequences

- If a sequence is a hybrid molecule; i.e., part of the backbone is DNA and part of the backbone is RNA:



DNA segment RNA segment

<u>ST.26 Paragraph 55 applies</u>:

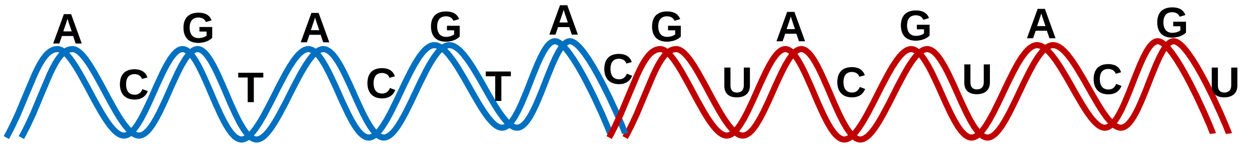
"55. For a nucleotide sequence that contains both DNA and RNA segments of one or more nucleotides, the molecule type must be indicated as DNA. The combined DNA/RNA molecule must be further described in the feature table, using the feature key "source" and the mandatory qualifier "organism" with the value "synthetic construct" and the mandatory qualifier "mol_type" with the value "other DNA". Each DNA and RNA segment of the combined DNA/RNA molecule must be further described with the feature key "misc_feature" and the qualifier "note", which indicates whether the segment is DNA or RNA."

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Special Situations

DNA/RNA Hybrid Sequences

- If a sequence is a hybrid molecule; i.e., part of the backbone is DNA and part of the backbone is RNA:



DNA segment

RNA segment

<u>ST.26 Paragraph 55 applies</u>:

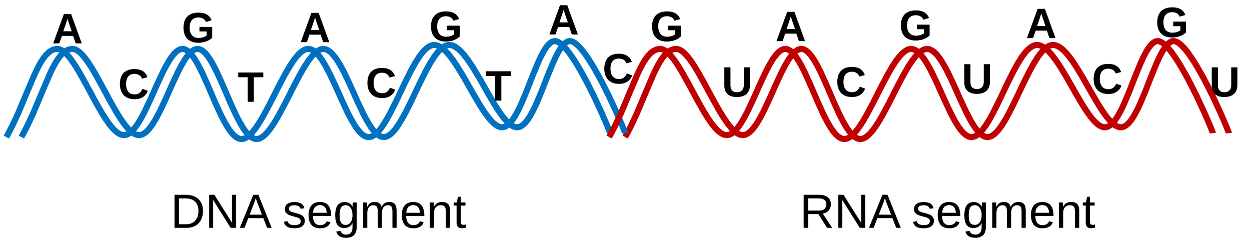
"55. For a nucleotide sequence that contains both DNA and RNA segments of one or more nucleotides, the molecule type must be indicated as DNA. The combined DNA/RNA molecule must be further described in the feature table, using the feature key "source" and the mandatory qualifier "organism" with the value "synthetic construct" and the mandatory qualifier "mol_type" with the value "other DNA". Each DNA and RNA segment of the combined DNA/RNA molecule must be further described with the feature key "misc_feature" and the qualifier "note", which indicates whether the segment is DNA or RNA."

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Special Situations

DNA/RNA Hybrid Sequences

- If a sequence is a hybrid molecule; i.e., part of the backbone is DNA and part of the backbone is RNA:



ST.26 Paragraph 55 applies:

"55. For a nucleotide sequence that contains both DNA and RNA segments of one or more nucleotides, the molecule type must be indicated as DNA. The combined DNA/RNA molecule must be further described in the feature table, using the feature key "source" and the mandatory qualifier "organism" with the value "synthetic construct" and the mandatory qualifier "mol_type" with the value "other DNA". Each DNA and RNA segment of the combined DNA/RNA molecule must be further described with the feature key "misc_feature" and the qualifier "note", which indicates whether the segment is DNA or RNA."

Special Situations

DNA/RNA Hybrid Sequences

An application discloses the following DNA/RNA hybrid sequence:

5′-ACCTGCcgucccacguguccgagguaGCATTA-3′

where upper-case symbols represent the DNA portion, and lower-case symbols represent the RNA portion.

- Residues 1-6 and 27-32 are DNA, residues 7-26 are RNA

- To consider:
  1. Organism designation
  2. molecule type and mol_type
  3. Identification of DNA and RNA segments

Special Situations

DNA and RNA Sequences

An application discloses the following DNA/RNA hybrid sequence:

5′-ACCTGCcgucccacguguccgagguaGCATTA-3′

where upper-case symbols represent the DNA portion, and lower-case symbols represent the RNA portion.

- ST.26 Paragraph 55 states: *“…the molecule type must be indicated as DNA. The combined DNA/RNA molecule must be further described in the feature table, using the feature key “source” and the mandatory qualifier “organism” with the value “synthetic construct” and the mandatory qualifier “mol_type” with the value “other DNA”.”*

✓ Molecule type = “DNA”

✓ Qualifier organism name = “synthetic construct”

✓ Qualifier mol_type = “other DNA”

Special Situations

DNA and RNA Sequences

An application discloses the following DNA/RNA hybrid sequence:

5′-ACCTGCcgucccacguguccgagguaGCATTA-3′

where upper-case symbols represent the DNA portion, and lower-case symbols represent the RNA portion.

✓ All uracil residues must be represented by the symbol “t”. Therefore, the sequence must be represented in the sequence listing as:

acctgccgtcccacgtgtccgaggtagcatta

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Special Situations

DNA and RNA Sequences

An application discloses the following DNA/RNA hybrid sequence:

```
5′-ACCTGCcgucccacguguccgagguaGCATTA-3′
```

where upper-case symbols represent the DNA portion, and lower-case symbols represent the RNA portion.

- ST.26 Paragraph 55 states: *“Each DNA and RNA segment of the combined DNA/RNA molecule must be further described with the feature key “misc_feature” and the qualifier “note”, which indicates whether the segment is DNA or RNA.”*

✓ Three segments = three “misc_feature” feature keys

Special Situations

DNA and RNA Sequences

An application discloses the following DNA/RNA hybrid sequence:

`5′-ACCTGCcgucccacguguccgagguaGCATTA-3′`

where upper-case symbols represent the DNA portion, and lower-case symbols represent the RNA portion.

✓ Segment 1, residues 1-6:
“misc_feature” feature key with location “1..6”
Qualifier “note” with value “DNA”

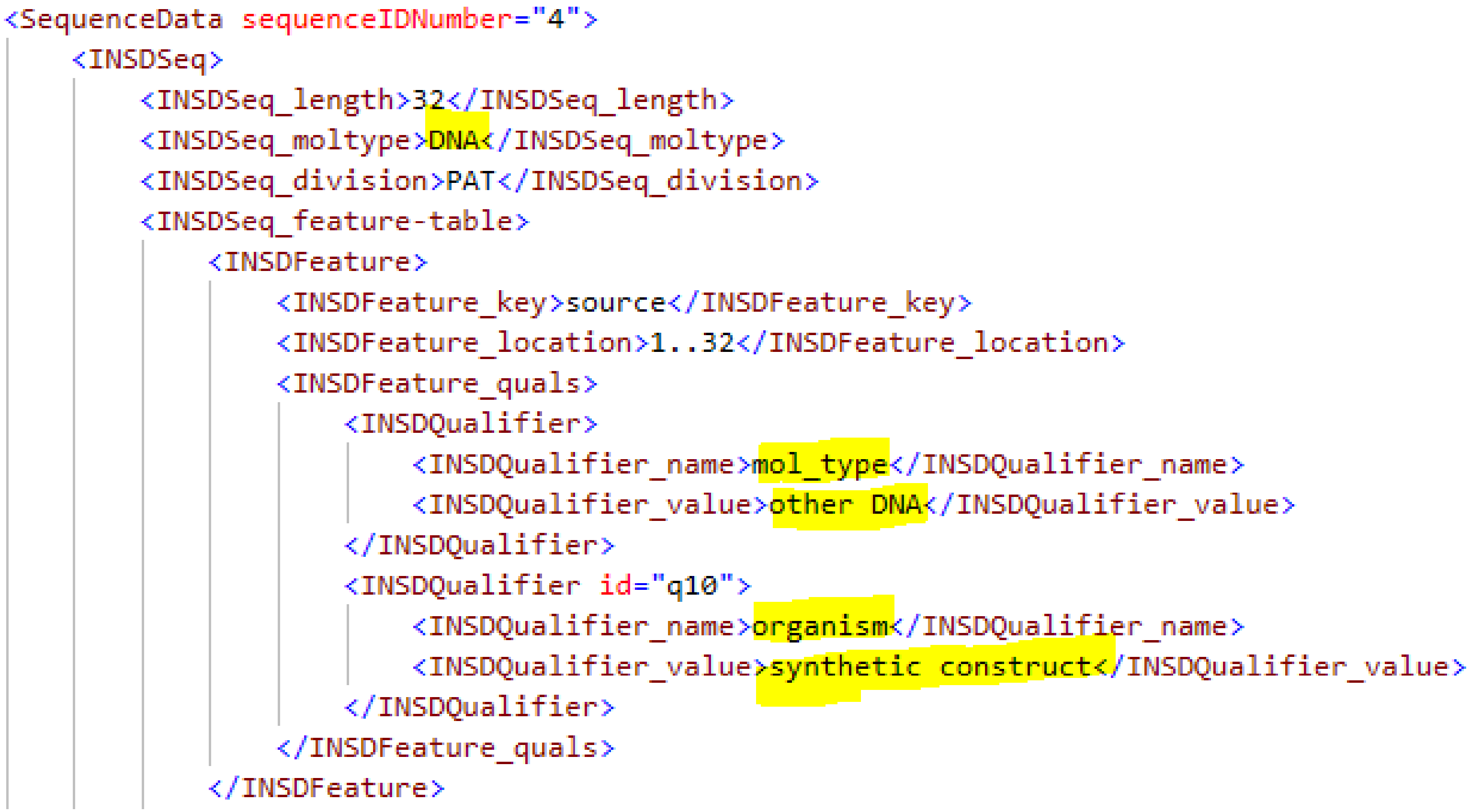
✓ Segment 2, residues 7-26:
“misc_feature” feature key with location “7..26”
Qualifier “note” with value “RNA”

✓ Segment 3, residues 27-32
“misc_feature” feature key with location “27..32”
Qualifier “note” with value “DNA”

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Special Situations

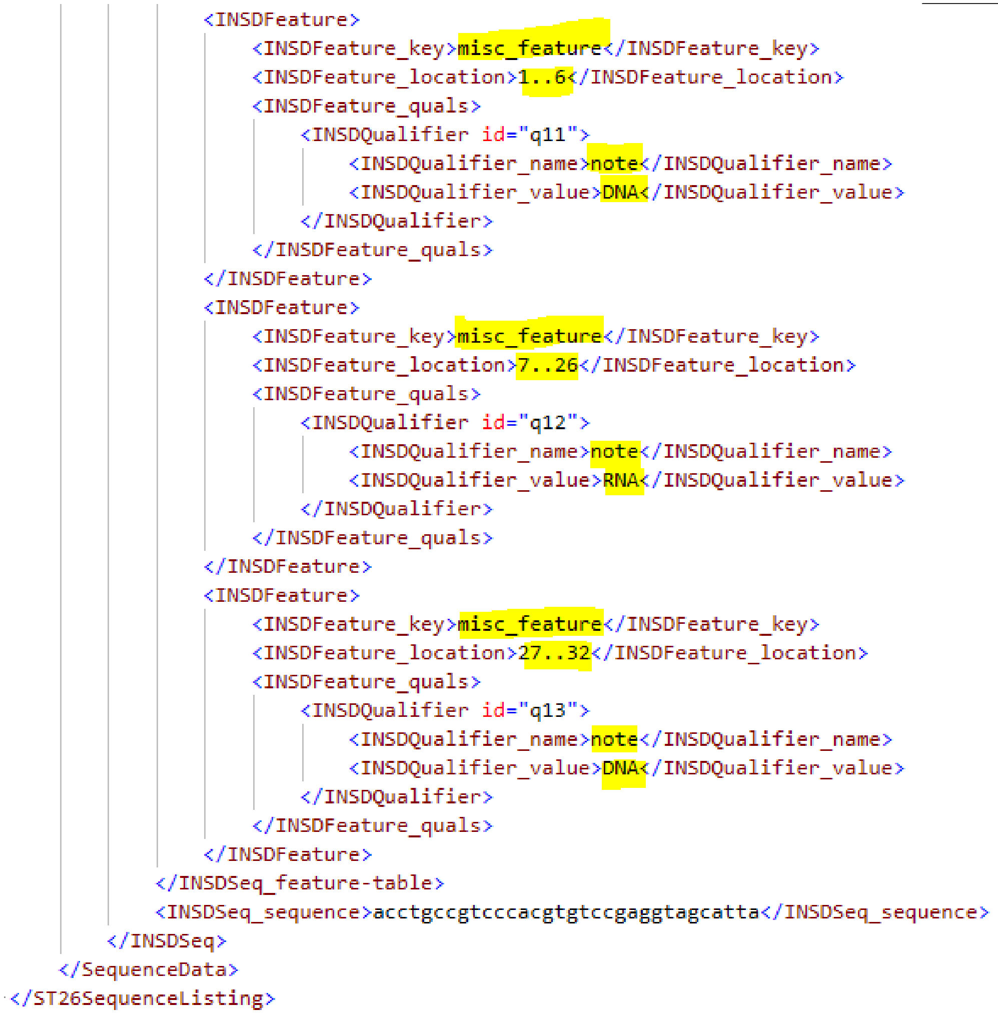
DNA and RNA Sequences



```
<SequenceData sequenceIDNumber="4">
    <INSDSeq>
        <INSDSeq_length>32</INSDSeq_length>
        <INSDSeq_moltype>DNA</INSDSeq_moltype>
        <INSDSeq_division>PAT</INSDSeq_division>
        <INSDSeq_feature-table>
            <INSDFeature>
                <INSDFeature_key>source</INSDFeature_key>
                <INSDFeature_location>1..32</INSDFeature_location>
                <INSDFeature_quals>
                    <INSDQualifier>
                        <INSDQualifier_name>mol_type</INSDQualifier_name>
                        <INSDQualifier_value>other DNA</INSDQualifier_value>
                    </INSDQualifier>
                    <INSDQualifier id="q10">
                        <INSDQualifier_name>organism</INSDQualifier_name>
                        <INSDQualifier_value>synthetic construct</INSDQualifier_value>
                    </INSDQualifier>
                </INSDFeature_quals>
            </INSDFeature>
```

Special Situations

DNA and RNA Sequences



```
        <INSDFeature>
          <INSDFeature_key>misc_feature</INSDFeature_key>
          <INSDFeature_location>1..6</INSDFeature_location>
          <INSDFeature_quals>
            <INSDQualifier id="q11">
              <INSDQualifier_name>note</INSDQualifier_name>
              <INSDQualifier_value>DNA</INSDQualifier_value>
            </INSDQualifier>
          </INSDFeature_quals>
        </INSDFeature>
        <INSDFeature>
          <INSDFeature_key>misc_feature</INSDFeature_key>
          <INSDFeature_location>7..26</INSDFeature_location>
          <INSDFeature_quals>
            <INSDQualifier id="q12">
              <INSDQualifier_name>note</INSDQualifier_name>
              <INSDQualifier_value>RNA</INSDQualifier_value>
            </INSDQualifier>
          </INSDFeature_quals>
        </INSDFeature>
        <INSDFeature>
          <INSDFeature_key>misc_feature</INSDFeature_key>
          <INSDFeature_location>27..32</INSDFeature_location>
          <INSDFeature_quals>
            <INSDQualifier id="q13">
              <INSDQualifier_name>note</INSDQualifier_name>
              <INSDQualifier_value>DNA</INSDQualifier_value>
            </INSDQualifier>
          </INSDFeature_quals>
        </INSDFeature>
      </INSDSeq_feature-table>
      <INSDSeq_sequence>acctgccgtcccacgtgtccgaggtagcatta</INSDSeq_sequence>
    </INSDSeq>
  </SequenceData>
</ST26SequenceListing>
```

DNA segment 1

RNA segment 2

DNA segment 3

Questions?

Nucleotide Analogs, D-Amino Acids, and Branched Sequences

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

New Molecule Types

Nucleotide Analogs

- Nucleic acid sequences that contain one or more nucleotide analogs are subject to ST.26 rules
- Nucleotide analogs are included in the definition of a “nucleotide” under ST.26, paragraph 3(g)(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”

- Common nucleotide analogs include peptide nucleic acids (PNAs), glycol nucleic acids (GNAs), threose nucleic acids, and morpholinos
- Must be represented in the direction from left to right that mimics the 5’ to 3’ direction. (ST.26, paragraph 11)

New Molecule Types

Nucleotide Analogs

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

PO_4-tagttcattgactaaggctccccattgact-OH

Wherein the PO_4 end of the sequence mimics the 5’ end of a DNA sequence.

✓ This sequence is required to be included in a sequence listing

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

New Molecule Types

Nucleotide Analogs

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

PO_4-tagttcattgactaaggctccccattgact-OH

Wherein the PO_4 end of the sequence mimics the 5’ end of a DNA sequence.

✓ This sequence is required to be included in a sequence listing

✓ The PO_4 end mimics the 5’ end, so the sequence must be represented in the direction shown

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

New Molecule Types

Nucleotide Analogs

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

PO_4-tagttcattgactaaggctccccattgact-OH

Wherein the PO_4 end of the sequence mimics the 5’ end of a DNA sequence.

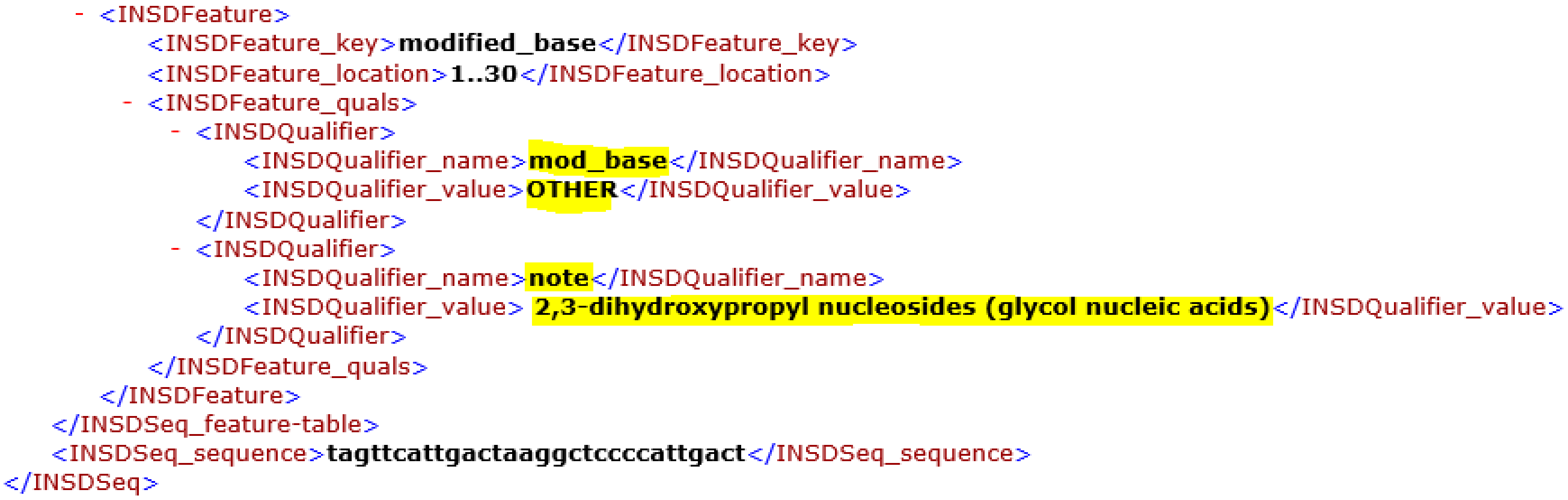
✓ This sequence is required to be included in a sequence listing

✓ The PO_4 end mimics the 5’ end, so the sequence must be represented in the direction shown

✓ The entire sequence must be annotated with the “modified_base” feature key, a “mod_base” qualifier with the value “OTHER”, and a note qualifier that includes the complete, unabbreviated name of the modified nucleotides, such as “glycol nucleic acids”

New Molecule Types

Nucleotide Analogs



```
 - <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>
```

Note: an extended discussion of this example can be found in WIPO Standard ST.26, Annex VI, Example 3(g)-4.

New Molecule Types

D-Amino Acids

- Amino acid sequences that contain one or more D-amino acids are subject to ST.26 rules
- D-amino acids are included in the definition of an “amino acid” under ST.26, paragraph 3(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.”

- D-amino acids must be represented in the sequence as the corresponding unmodified L amino acid symbol, where possible

- must be described in the feature table as a modified amino acid

New Molecule Types

D-Amino Acids

A patent application describes the following sequence:

D-Ala-D-Glu-Lys-Leu-Gly-D-Met

✓ This sequence is required to be included in a sequence listing

New Molecule Types

D-Amino Acids

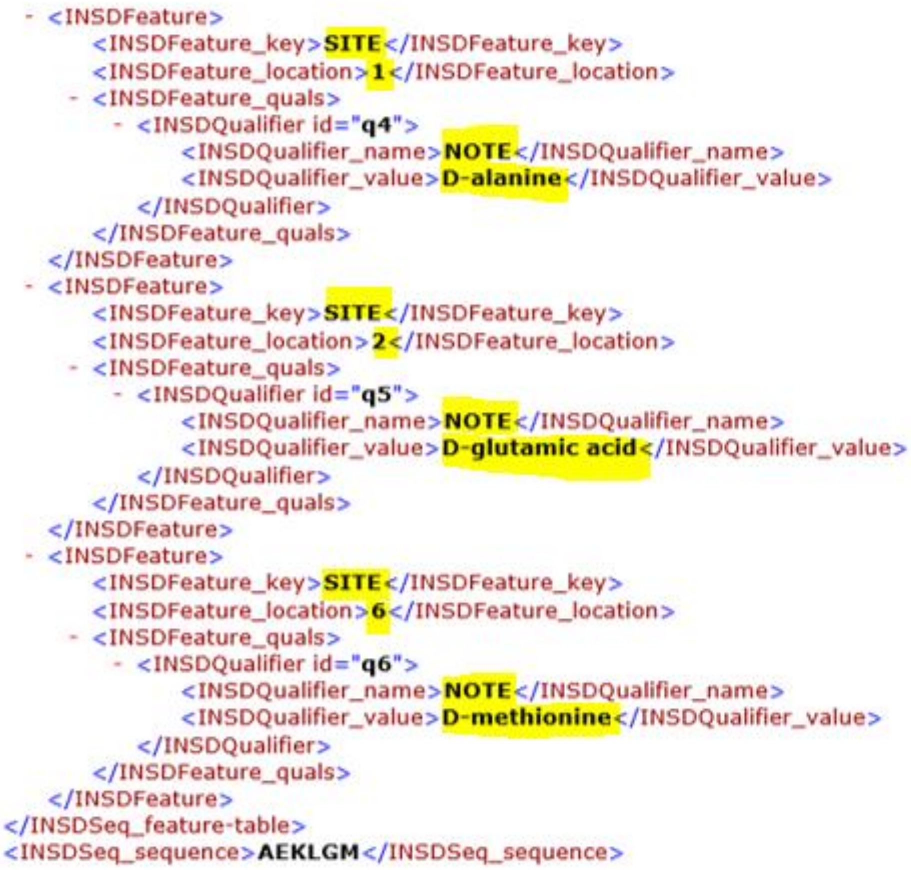
A patent application describes the following sequence:

D-Ala-D-Glu-Lys-Leu-Gly-D-Met

- ✓ This sequence is required to be included in a sequence listing
- ✓ Should be represented as: AGKLGM
- ✓ The alanine in position 1, the glutamic acid in position 2, and the methionine in position 6 must each be annotated with a “SITE” feature key and a “NOTE” qualifier with the complete, unabbreviated name of the corresponding amino acid

New Molecule Types

D-Amino Acids



```
- <INSDFeature>
    <INSDFeature_key>SITE</INSDFeature_key>
    <INSDFeature_location>1</INSDFeature_location>
  - <INSDFeature_quals>
    - <INSDQualifier id="q4">
        <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 id="q5">
        <INSDQualifier_name>NOTE</INSDQualifier_name>
        <INSDQualifier_value>D-glutamic acid</INSDQualifier_value>
      </INSDQualifier>
    </INSDFeature_quals>
  </INSDFeature>
- <INSDFeature>
    <INSDFeature_key>SITE</INSDFeature_key>
    <INSDFeature_location>6</INSDFeature_location>
  - <INSDFeature_quals>
    - <INSDQualifier id="q6">
        <INSDQualifier_name>NOTE</INSDQualifier_name>
        <INSDQualifier_value>D-methionine</INSDQualifier_value>
      </INSDQualifier>
    </INSDFeature_quals>
  </INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>AEKLGM</INSDSeq_sequence>
```

Note: an extended discussion of a similar example can be found in WIPO Standard ST.26, Annex VI, Example 3(a)-1.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

New Molecule Types

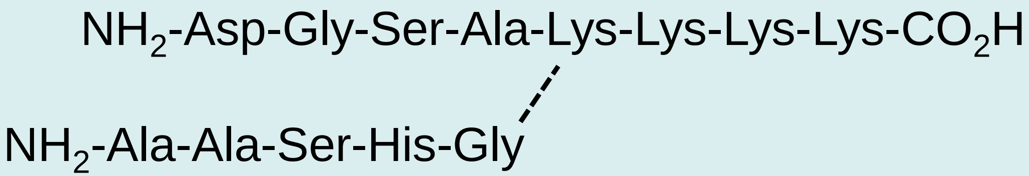
Branched Sequences

- Branched nucleic acid sequences and branched amino acid sequences are subject to ST.26 rules
- Linear regions of branched sequences must be included in a sequence listing when they contain 10 or more specifically defined nucleotides or 4 or more specifically defined amino acids. (WIPO Standard ST.26, paragraph 7)
- Each linear region of a branched sequence that meets the minimum length requirement must be included as a separate sequence with its own SEQ ID number
- The number of specifically defined residues in each individual linear region must be considered, not the total number of specifically defined residues in the structure

New Molecule Types

Branched Sequences

A patent application describes a peptide with the following sequence:



NH_2-Asp-Gly-Ser-Ala-Lys-Lys-Lys-Lys-CO_2H

NH_2-Ala-Ala-Ser-His-Gly

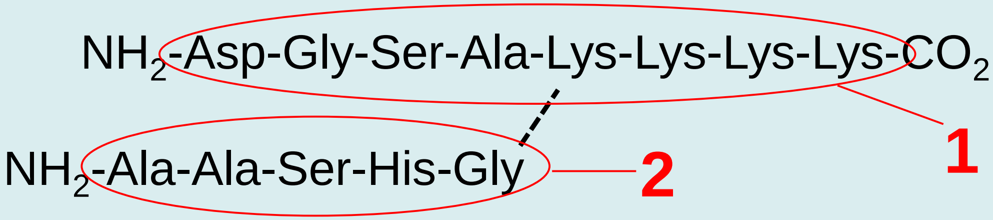
Where ----- indicates an amide bond between the carboxy terminus of the glycine and the side chain of the lysine

✓ Both linear regions contain ≥ 4 specifically defined amino acids, so both are required to be included in a sequence listing

New Molecule Types

Branched Sequences

A patent application describes a peptide with the following sequence:



NH_2-Asp-Gly-Ser-Ala-Lys-Lys-Lys-Lys-CO_2H

NH_2-Ala-Ala-Ser-His-Gly

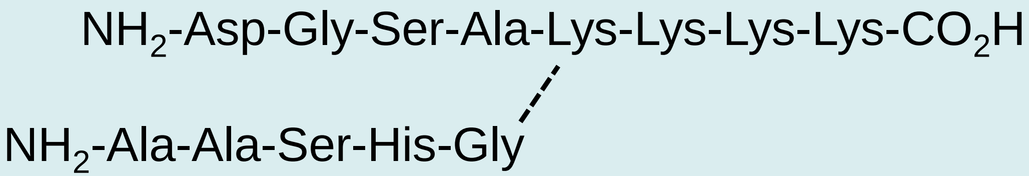
Where ----- indicates an amide bond between the carboxy terminus of the glycine and the side chain of the lysine

- ✓ Both linear regions contain ≥ 4 specifically defined amino acids, so both are required to be included in a sequence listing
- ✓ Each linear region must be included as a separate sequence with its own SEQ ID number

New Molecule Types

Branched Sequences

A patent application describes a peptide with the following sequence:



NH_2-Asp-Gly-Ser-Ala-Lys-Lys-Lys-Lys-CO_2H

NH_2-Ala-Ala-Ser-His-Gly

Where ----- indicates an amide bond between the carboxy terminus of the glycine and the side chain of the lysine

✓ Both linear regions contain ≥ 4 specifically defined amino acids, so both are required to be included in a sequence listing

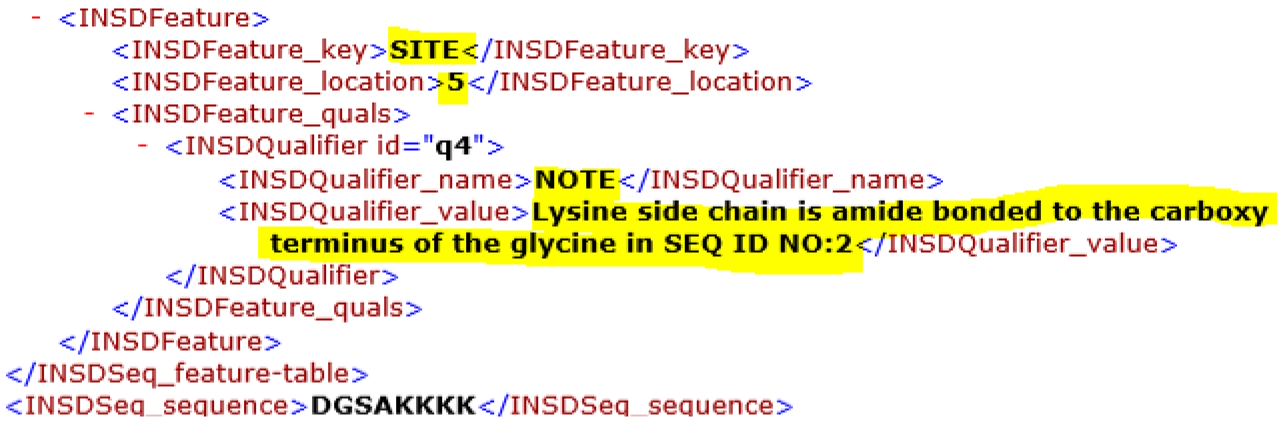
✓ Each linear region must be included as a separate sequence with its own SEQ ID number

✓ Both sequences should be annotated to indicate the location and nature of the amide bond linkage

New Molecule Types

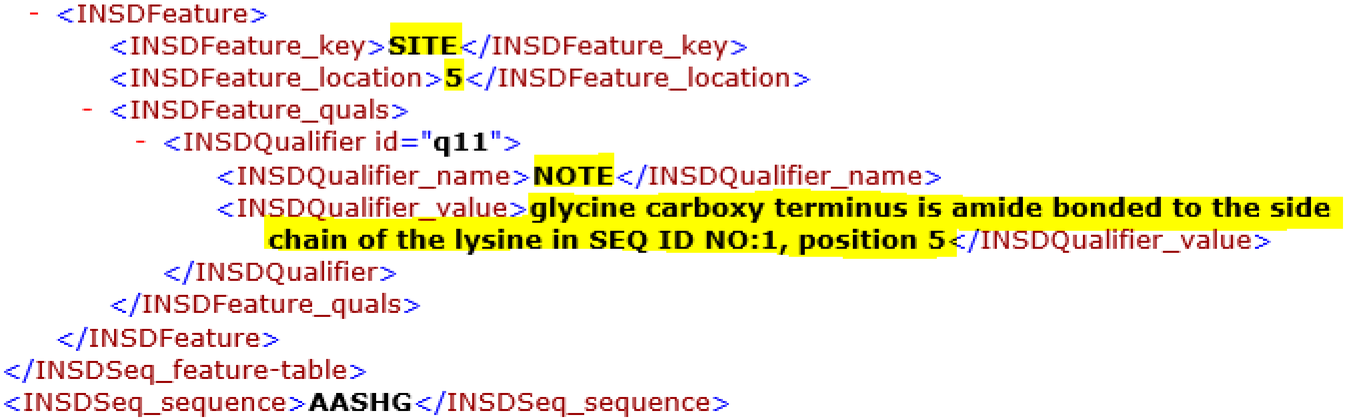
Branched Sequences

1



```
  - <INSDFeature>
        <INSDFeature_key>SITE</INSDFeature_key>
        <INSDFeature_location>5</INSDFeature_location>
      - <INSDFeature_quals>
          - <INSDQualifier id="q4">
                <INSDQualifier_name>NOTE</INSDQualifier_name>
                <INSDQualifier_value>Lysine side chain is amide bonded to the carboxy
                  terminus of the glycine in SEQ ID NO:2</INSDQualifier_value>
            </INSDQualifier>
        </INSDFeature_quals>
    </INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>DGSAKKKK</INSDSeq_sequence>
```

2



```
  - <INSDFeature>
        <INSDFeature_key>SITE</INSDFeature_key>
        <INSDFeature_location>5</INSDFeature_location>
      - <INSDFeature_quals>
          - <INSDQualifier id="q11">
                <INSDQualifier_name>NOTE</INSDQualifier_name>
                <INSDQualifier_value>glycine carboxy terminus is amide bonded to the side
                  chain of the lysine in SEQ ID NO:1, position 5</INSDQualifier_value>
            </INSDQualifier>
        </INSDFeature_quals>
    </INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>AASHG</INSDSeq_sequence>
```

Note: an extended discussion of a similar example can be found in WIPO Standard ST.26, Annex VI, Example 7(b)-3.

Sequence Variants

Sequence Variants

- ST.26 paragraph 3(m) defines a “***variant sequence***” as

*“a nucleotide or amino acid sequence that contains **one or more differences with respect to a primary sequence**. These differences may include alternative residues (see paragraphs 15 and 27), modified residues (see paragraphs 3(g), 3(h), 16, and 29), deletions, insertions, and substitutions. See paragraphs 93 to 95*.”

- The manner in which a variant sequence is disclosed will determine how it must be represented in a sequence listing.

- ST.26 paragraphs 93-95 control how variants must be represented.

Sequence Variants

Paragraph 93

Paragraph 93: A primary sequence and any variant of that sequence, <u>each disclosed by enumeration of their residues</u> and encompassed by paragraph 7, must each be included in the sequence listing and assigned their own sequence identification number.

If each variant is enumerated separately, then each variant <u>must</u> have its own SEQ ID number!

Sequence Variants

Paragraph 93

A patent application includes a figure with the following multiple sequence alignment:

Consensus	LEG**n**EQFINA**ak**IIRHP**k**Y**nrk**Tl**n**NDI**m**LIK
Homo sapiens	LEGNEQFINAAKIIRHPQYDRKTLNNDIMLIK
Pongo abelii	LEGNEQFINAAKIIRHPQYDRKTVNNDIMLIK
Pan paniscus	LEGNEQFINAAKIIRHPKYNRITLNNDIMLIK
Rhinopithecus bieti	LEGNEQFINATKIIRHPKYNGNTLNNDIMLIK
Rhinopithecus roxellana	LEGNEQFINATQIIRHPKYNGNTLNNDIMLIK

Lower case letters represent the predominant amino acid residues among the aligned sequences.

✓ Each of the 6 enumerated sequences must be included in the sequence listing as a separate sequence with its own SEQ ID number.

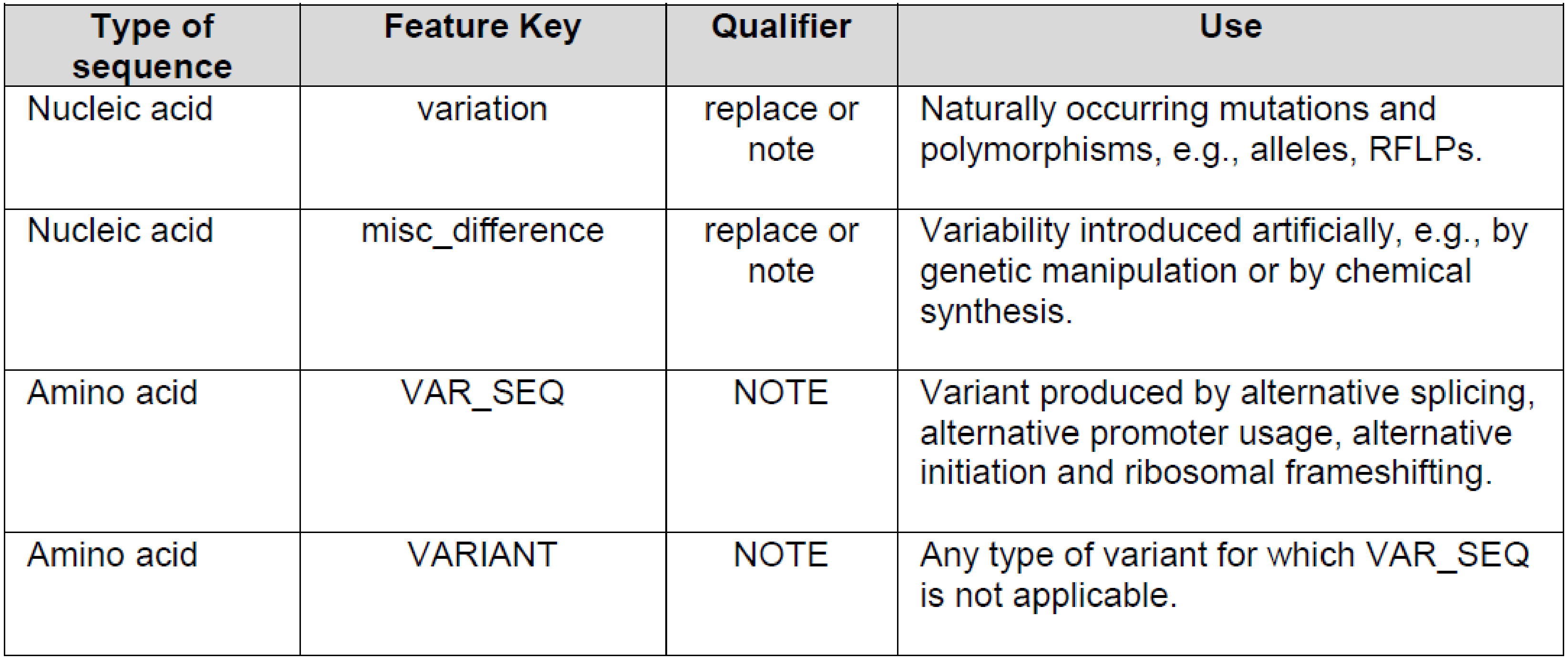
Note: an extended discussion of a similar example can be found in WIPO Standard ST.26, Annex VI, Example 93-3.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Sequence Variants

Proper Annotation

- The proper feature key to annotate a variant depends on the molecule type and the nature of the variation:



Type of sequence	Feature Key	Qualifier	Use
Nucleic acid	variation	replace or note	Naturally occurring mutations and polymorphisms, e.g., alleles, RFLPs.
Nucleic acid	misc_difference	replace or note	Variability introduced artificially, e.g., by genetic manipulation or by chemical synthesis.
Amino acid	VAR_SEQ	NOTE	Variant produced by alternative splicing, alternative promoter usage, alternative initiation and ribosomal frameshifting.
Amino acid	VARIANT	NOTE	Any type of variant for which VAR_SEQ is not applicable.

WIPO Standard ST.26, paragraph 96

Sequence Variants

Most Restrictive Ambiguity Symbol

- ST.26 paragraphs 15 and 27 state that when an ambiguity symbol is required, "the most restrictive symbol should be used…"

  What does that mean?

For example:

In a nucleotide sequence, if a position can be "a or c", use the ambiguity symbol "m" instead of "n".

In an amino acid sequence, if a position can be "L or I", use the ambiguity symbol "J" instead of "X".

Remember that "n" and "X" have default values, so any time "n" or "X" are used for something other than the default value, an annotation is required.

Sequence Variants

Paragraph 94

Paragraph 94: Any variant sequence, disclosed as a single sequence with enumerated alternative residues at one or more positions, must be included in the sequence listing and should be represented by a single sequence, wherein the enumerated alternative residues are represented by the most restrictive ambiguity symbol (see paragraphs 15 and 27).

If variants ARE NOT separately enumerated, but are simply shown as variable residues in the primary sequence, then they are not required to have a separate SEQ ID number!

Sequence Variants

Paragraph 94

A patent application discloses a peptide of the sequence:

Gly-Gly-Gly-[Leu or Ile]-Ala-Thr-[Ser or Thr]

- ✓ May be included in the sequence listing as a single sequence
- ✓ The preferred representation of the sequence is: GGGJATX
- ✓ [Leu or Ile] should be represented by the most restrictive ambiguity symbol “J”
- ✓ [Ser or Thr] should be represented by the symbol “X” along with the feature key “VARIANT” with a qualifier note that indicates that X is serine or threonine

Note: an extended discussion of this example can be found in WIPO Standard ST.26, Annex VI, Example 94-1.

Sequence Variants
Paragraph 95

Paragraph 95: Any variant sequence, disclosed only by reference to deletion(s), insertion(s), or substitution(s) in a primary sequence in the sequence listing, should be included in the sequence listing. Where included in the sequence listing, such a variant sequence:

(a) may be represented by annotation of the primary sequence, where it contains variation(s) at a single location or multiple distinct locations and the occurrence of those variations are independent;

(b) should be represented as a separate sequence and assigned its own sequence identification number, where it contains variations at multiple distinct locations and the occurrence of those variations are interdependent; and

(c) must be represented as a separate sequence and assigned its own sequence identification number, where it contains an inserted or substituted sequence that contains in excess of 1000 residues (see paragraph 86).

"Reference to deletion(s), insertion(s), or substitution(s)" means the variants are disclosed in prose.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Sequence Variants

Paragraph 95(a)

An application contains the following disclosure:

Peptide fragment 1: Gly-Leu-Pro-Xaa-Arg-Ile-Cys
wherein Xaa is 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…

The primary sequence, Gly-Leu-Pro-Xaa-Arg-Ile-Cys, *contains variation(s) at a single location and the occurrence of those variations are independent*

Sequence Variants

Paragraph 95(a)

✓ The primary sequence must be included in the sequence listing, and variants may be represented by annotation of the primary sequence

✓ The most encompassing embodiment must be the version included in the sequence listing – in this example, it is the version where Xaa is “any amino acid”

✓ The sequence must be represented as GLPXRIC and requires the feature key “VARIANT” at position 4 and a qualifier “note” indicating that Xaa is “any amino acid”

✓ While not required, it is recommended that the following three variants are included in the sequence listing as separate sequences:

GLPVRIC

GLPTRIC

GLPDRIC

Note: an extended discussion of this example can be found in WIPO Standard ST.26, Annex VI, Example 95(a)-1.

Sequence Variants

Paragraph 95(b)

A patent application describes the following consensus sequence:

aatg$\mathbf{n_1}$cccacgaatg$\mathbf{n_2}$cac

wherein $\mathbf{n_1}$ and $\mathbf{n_2}$ can be a, t, g, or c.

Several variant sequences are disclosed as follows:

if n_1 is a, then n_2 is t, g, or c
if n_1 is t, then n_2 is a, g, or c
if n_1 is g, then n_2 is t, a, or c
if n_1 is c, then n_2 is t, g, or a

The consensus sequence contains variations at multiple distinct locations and the occurrence of those variations are interdependent

Sequence Variants

Paragraph 95(b)

✓ The consensus sequence must be included in the sequence listing, and variants should be represented as separate sequences

✓ The most encompassing embodiment must be the version included in the sequence listing: where $\mathbf{n}_1$ and $\mathbf{n}_2$ “can be a, t, g, or c”

✓ The sequence must be represented as: aatg**n**cccacgaatg**n**cac

✓ No annotation for n is required, since "n” is interpreted as one of "a", "c", "g" or "t" in the absence of a feature key (see ST.26 paragraph 15)

✓ While not required, it is strongly recommended that the following four variants are included in the sequence listing as separate sequences:

aatgacccacgaatgbcac	(b = t, g, or c)
aatgtcccacgaatgvcac	(v = a, g, or c)
aatggcccacgaatghcac	(h = t, a, or c)
aatgccccacgaatgdcac	(d = t, g, or a)

Note: an extended discussion of a similar example can be found in WIPO Standard ST.26, Annex VI, Example 95(b).

Sequence Variants

Paragraph 95(c)

An application contains the following disclosure:

… -Met-Gly-Leu-Pro-Arg-Xaa-Arg-Ile-Cys-Lys- …

wherein Xaa is Gly or an insertion of the sequence

Cys-Tyr-Ile-Lys-Ser-(1000 amino acids)-Leu-Thr-Pro-Lys

One variant sequence contains an inserted or substituted sequence in excess of 1000 residues

Sequence Variants

Paragraph 95(c)

- ✓ The variant where Xaa = an insertion of over 1000 residues must be included in the sequence listing as a separate sequence with its own SEQ ID number.

- ✓ The variant where Xaa = Gly will also be included in the sequence listing as a separate sequence with its own SEQ ID number.

…-MGLPRGRICK-…

Sequence Variants

Q: Which paragraph applies to the following disclosure?

G-L-P-T-R-I-C-[L or I]-A-V-[G or A]

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Sequence Variants

Q: Which paragraph applies to the following disclosure?

```
G-L-P-T-R-I-C-[L or I]-A-V-[G or A]
```

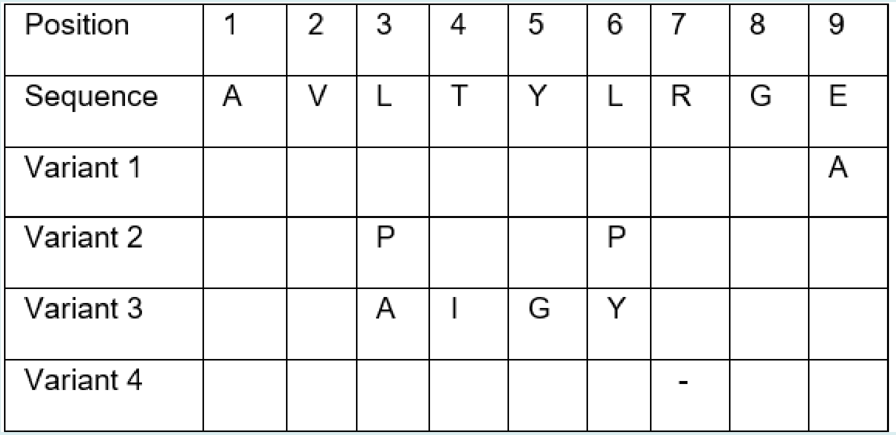
A: Paragraph 94:

“*Any variant sequence, <u>disclosed as a single sequence with enumerated alternative residues at one or more positions</u>, must be included in the sequence listing and should be represented by a single sequence, wherein the enumerated alternative residues are represented by the most restrictive ambiguity symbol.*”

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Sequence Variants

Q: Which paragraph applies to the following disclosure?

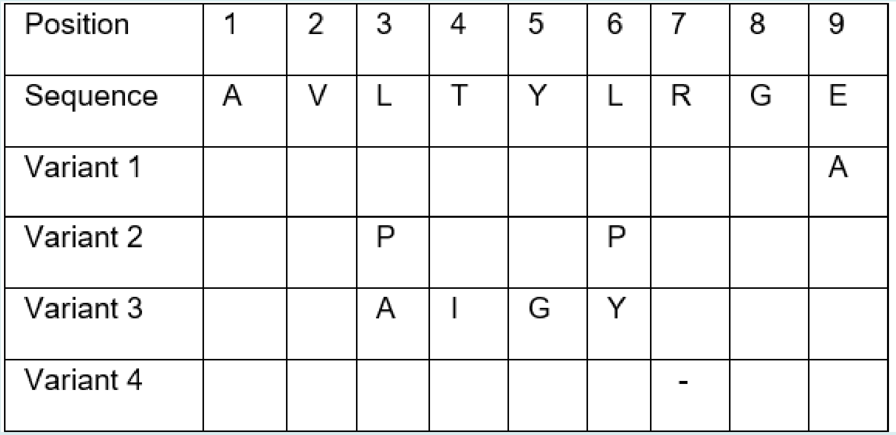


Position	1	2	3	4	5	6	7	8	9
Sequence	A	V	L	T	Y	L	R	G	E
Variant 1									A
Variant 2			P			P			
Variant 3			A	I	G	Y			
Variant 4							-		

A blank space in the table indicates that an amino acid in the variant is the same as the corresponding amino acid in the “Sequence” and a “-“ indicates deletion of the corresponding amino acid in the “Sequence”.

Sequence Variants

Q: Which paragraph applies to the following disclosure?



Position	1	2	3	4	5	6	7	8	9
Sequence	A	V	L	T	Y	L	R	G	E
Variant 1									A
Variant 2			P			P			
Variant 3			A	I	G	Y			
Variant 4							-		

A blank space in the table indicates that an amino acid in the variant is the same as the corresponding amino acid in the "Sequence" and a "-" indicates deletion of the corresponding amino acid in the "Sequence".

A: Paragraph 93

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

Sequence Variants

Q: Which paragraph applies to the following disclosure?

A figure discloses the following sequence:

Met-Gly-Ala-Ile-Pro-Asp-Val-Lys-Arg-Ala-Cys-Trp (Sequence 1)

The specification contains the following information concerning Sequence 1:

… in certain embodiments, the valine at position 7 of sequence 1 is replaced with alanine…

Sequence Variants

Q: Which paragraph applies to the following disclosure?

A figure discloses the following sequence:

Met-Gly-Ala-Ile-Pro-Asp-Val-Lys-Arg-Ala-Cys-Trp (Sequence 1)

The specification contains the following information concerning Sequence 1:

… in certain embodiments, the valine at position 7 of sequence 1 is replaced with alanine…

A: Paragraph 95(a)*: "Any variant sequence, disclosed only by reference to deletion(s), insertion(s), or substitution(s) in a primary sequence in the sequence listing, should be included in the sequence listing. Where included in the sequence listing, such a variant sequence:*

(a) may be represented by annotation of the primary sequence, where it contains variation(s) at a single location or multiple distinct locations and the occurrence of those variations are independent;"

Sequence Variants

Q: Which paragraph applies to the following disclosure?

A figure discloses the following sequence:

Met-Gly-Ala-Ile-Pro-Asp-Val-Lys-Arg-Ala-Cys-Trp (Sequence 1)

The specification contains the following information concerning Sequence 1:

… if the valine at position 7 of sequence 1 is replaced with alanine then the alanine at position 10 is replaced with valine….

Sequence Variants

Q: Which paragraph applies to the following disclosure?

A figure discloses the following sequence:

Met-Gly-Ala-Ile-Pro-Asp-Val-Lys-Arg-Ala-Cys-Trp (Sequence 1)

The specification contains the following information concerning Sequence 1:

… if the valine at position 7 of sequence 1 is replaced with alanine then the alanine at position 10 is replaced with valine….

A: Paragraph 95(b)*: Any variant sequence, <u>disclosed only by reference to deletion(s), insertion(s), or substitution(s)</u> in a primary sequence in the sequence listing, <u>should</u> be included in the sequence listing. Where included in the sequence listing, such a variant sequence:*

(b) should be represented as a separate sequence and assigned its own sequence identification number, where it contains variations at multiple distinct locations and the occurrence of those <u>variations are interdependent</u>;

Questions?