

WIPO标准ST.26 进阶

网络研讨会培训

注：本培训为进阶模块，预设参与者已参加过WIPO ST.26入门培训模块

可由此获取WIPO ST.26入门课程记录：https://www.wipo.int/meetings/en/details.jsp?meeting_id=62848

培训内容

- 常用特征键和限定符
- 特征定位格式
- 限定符值格式和非英语语言限定符值
- 特殊情况——DNA中的尿嘧啶和RNA中的胸腺嘧啶；DNA/RNA杂交分子
- 核苷酸类似物、D-氨基酸和分支序列
- 序列变异

常用特征键和限定符

特征键和限定符

- 特征键可以用来进一步描述以位置标识的序列的一个或多个残基
 - 核苷酸序列的特征键列于附件一第**5**部分
 - 核苷酸序列的特征键是小写；例如，
“misc_binding”
 - 氨基酸序列的特征键列于附件一第**7**部分
 - 氨基酸序列的特征键是大写；例如，
“REGION”

- 限定符可以用来进一步描述特征
 - 核苷酸序列的限定符列于附件一第**6**部分
 - 核苷酸序列的限定符是小写；例如，
“allele”
 - 氨基酸序列的限定符列于附件一第**8**部分
 - 氨基酸序列的限定符是大写；例如，
“NOTE”

特征键和限定符

```

<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_qual>
    - <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_qual>
  </INSDFeature>
- <INSDFeature>
  <INSDFeature_key>VARIANT</INSDFeature_key>
  <INSDFeature_location>1</INSDFeature_location>
- <INSDFeature_qual>
  - <INSDQualifier>
    <INSDQualifier_name>NOTE</INSDQualifier_name>
    <INSDQualifier_value>X can be any amino acid</INSDQualifier_value>
  </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>
  <INSDSeq_sequence>XYEKGJL</INSDSeq_sequence>
</INSDSeq>
</SequenceData>

```

特征键和限定符

- 特征键为可选项，只有“source”/“SOURCE”是每个序列的强制性特征。
- 每个特征键都有一系列限定符，可用来进一步描述特征。多数限定符是可选项；然而有些特征键具有强制限定符。

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

- “mol_type”/“MOL_TYPE”和“organism”/“ORGANISM”是“source”/“SOURCE”特征的强制限定符。

特征键和限定符

- 有些特征键有额外限制
 - 有机体范围；例如，“C_region”仅限于真核生物
 - 分子范围；例如，“D-loop”仅限于DNA序列

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

特征键和限定符

核苷酸序列：“modified_base”特征键

- 特征键“modified_base”及其强制定限定符“mod_base”应用来描述改性核苷酸（ST.26，第16段）
- “改性核苷酸”是除以下外所有核苷酸：
 - 脱氧-[a, g, c, or t] 3'-单磷酸盐
 - [a, g, c, or u] 3'-单磷酸盐（ST.26，第3段（f））
- 在可能的情况下，“改性核苷酸”应以对应的未改性核苷酸表示（附件一第1部分表1）。否则，可以用“n”来表示。例如，在序列中，“2'-O-甲基胞苷”应以“c”来表示。“嘌呤”应以“n”来表示。符号“n”仅等于一个残基。
- 强制定限定符“mod_base”的值必须选自附件一第2部分表2。如果值为“other”，则额外的“note”限定符中必须包含改性残基的完整非缩略名称。

特征键和限定符

核苷酸序列：“modified_base”特征键

- 例如：位置15为肌昔的核苷酸序列

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

- “肌昔”在附件一第2部分表2中以缩写“i”列出

特征键和限定符

核苷酸序列：“modified_base”特征键

- 例如：位置22为黄嘌呤的核苷酸序列

```
<INSDFeature>
  <INSDFeature_key>modified_base</INSDFeature_key>
  <INSDFeature_location>22</INSDFeature_location>
  <INSDFeature_qual>
    <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_qual>
</INSDFeature>
```

- 附件一第2部分表2中并未列出“黄嘌呤”；因此，“mod_base”强制定限定符的值必须是“OTHER”，此外还必须加入值为“xanthine”的“note”限定符。

特征键和限定符

核苷酸序列：“modified_base”特征键

- “modified_base”也可以用来描述无碱基位点：

```

<INSDFeature>
  <INSDFeature_key>modified_base</INSDFeature_key>
  <INSDFeature_location>11</INSDFeature_location>
  <INSDFeature_qual>
    <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_qual>
</INSDFeature>
  
```

- 在序列中，可以用“n”代表无碱基位点，并用“modified_base”特征键以及值为“OTHER”的“mod_base”限定符和值为“abasic site”的“note”额外限定符进一步描述。

特征键和限定符

核苷酸序列：“CDS”特征键

- “CDS”特征键可被用于标识编码序列。CDS特征定位必须包括终止密码子。
(ST.26, 第89段)
- “CDS”特征键没有强制限定符
- 常用限定符包括：

“pseudo”

“pseudogene”

“translation”

“transl_table”

“codon_start”

“transl_except”

“protein_id”

特征键和限定符

核苷酸序列：“CDS”特征键

- “CDS”特征只能有以下限定符中的一个：

限定符	描述	值
pseudo	表示CDS特征不具功能性，且没有翻译，但并非伪基因	无
pseudogene	表示CDS特征是伪基因，并且没有翻译	已处理 已处理 单一 等位 未知
translation	表明是来自翻译CDS的氨基酸序列	单字母氨基酸缩写

特征键和限定符

核苷酸序列：“CDS”特征键

- 包含四个或更多特定氨基酸的氨基酸序列，如果是由编码序列进行编码并且在“**translation**”限定符中进行公开，就必须作为单独序列纳入序列表，并且必须为其分配单独的序列标识编号。
- 分配给氨基酸序列的序列标识编号必须是“CDS”特征键中“**protein_id**”限定符的值。
- 氨基酸序列“**SOURCE**”特征键的“**ORGANISM**”限定符必须与其编码序列中的相同。（ST.26，第92段）

特征键和限定符

核苷酸序列：“CDS”特征键

- 能够修改已翻译序列的限定符：

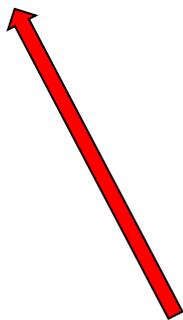
限定符	描述	值
transl_table	表示用来翻译CDS的基因编码表；默认为“1-Standard Code”	与附件一第9部分中翻译表相符的数字
transl_except	表示与“transl_table”中所确定的基因编码不符的密码子的翻译	(pos:<location>, aa:<amino_acid>)
codon_start	表示CDS相对于1位碱基的阅读框架	1、2或3

特征键和限定符

核苷酸序列：“CDS”特征键

- 例如——SEQ ID NO:1是来自酵母线粒体基因的编码序列片段（基因编码表“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 *
```



序列开头为部分密码子



在位置30-32的密码子

对硒半胱氨酸（Sec）进行编码

- 为准确表示这一序列，CDS特征中应包含哪些信息？

特征键和限定符

核苷酸序列：“CDS”特征键

- 例如——SEQ ID NO:1是来自酵母线粒体基因的编码序列片段（基因编码表“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 *
```

- 序列表中呈现的序列为：

```
tggataatga agaagttaac gaagaatgta tgagattatt ttcaagaac gctcgtcatc taacatcaag gttgacataa
```

特征键和限定符

核苷酸序列：“CDS”特征键

- 例如——SEQ ID NO:1是来自酵母线粒体基因的编码序列片段（基因编码表“3-酵母线粒体编码”）：

```
...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”特征键
- 特征定位：<1..80

“<” 1表示编码区始于位置1之前

定位包括在位置78-80的终止密码子

特征键和限定符

核苷酸序列：“CDS”特征键

- 例如——SEQ ID NO:1是来自酵母线粒体基因的编码序列片段（基因编码表“3-酵母线粒体编码”）：

```
...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”特征键
- 特征定位：<1..80
- 值为“3”的“codon_start”限定符

值为3的“codon_start”限定符
表示第一个完整密码子始自
定位中的第3个位置

特征键和限定符

核苷酸序列：“CDS”特征键

- 例如——SEQ ID NO:1是来自酵母线粒体基因的编码序列片段（基因编码表“3-酵母线粒体编码”）：

```
...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”特征键
- 特征定位：<1..80
- 值为“3”的“codon_start”限定符
- 值为“3”的“transl_table”限定符

使用附件一第9部分中的
基因编码表来确定
“transl_table”限定符的
值

特征键和限定符

核苷酸序列：“CDS”特征键

- 例如——SEQ ID NO:1是来自酵母线粒体基因的编码序列片段（基因编码表“3-酵母线粒体编码”）：

```
...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”特征键
- 特征定位：<1..80
- 值为“3”的“codon_start”限定符
- 值为3的“transl_table”限定符
- 值为“(pos:30..32,aa:Sec)”的“transl_except”限定符

“transl_except”限定符将补充
氨基酸硒半胱氨酸

特征键和限定符

核苷酸序列：“CDS”特征键

- 例如——SEQ ID NO:1是来自酵母线粒体基因的编码序列片段（基因编码表“3-酵母线粒体编码”）：

```
...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”特征键
- 特征定位：<1..80
- 值为“3”的“codon_start”限定符
- 值为3的“transl_table”限定符
- 值为“(pos:30..32,aa:Sec)”的“transl_except”限定符
- 值为“**DNEEVNEECMRLFFKNARHTTSRLT**”的“translation”限定符

在翻译限定符中不显示
终止密码子！

特征键和限定符

核苷酸序列：“CDS”特征键

- 例如——SEQ ID NO:1是来自酵母线粒体基因的编码序列片段（基因编码表“3-酵母线粒体编码”）：

```
...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”特征键
- 特征定位：<1..80
- 值为“3”的“codon_start”限定符
- 值为3的“transl_table”限定符
- 值为“(pos:30..32,aa:Sec)”的“transl_except”限定符
- 值为“**DNEEVNEECMRLFFKNARHTTSRLT**”的“translation”限定符
- 供翻译的单独蛋白质序列

特征键和限定符

核苷酸序列：“CDS”特征键

- 例如——SEQ ID NO:1是来自酵母线粒体基因的编码序列片段（基因编码表“3-酵母线粒体编码”）：

```
...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”特征键
- 特征定位：<1..80
- 值为“3”的“codon_start”限定符
- 值为3的“transl_table”限定符
- 值为“(pos:30..32,aa:Sec)”的“transl_except”限定符
- 值为“**DNEEVNEECMRLFFKNARHTTSRLT**”的“translation”限定符
- 供翻译的单独蛋白质序列
- 带有已翻译蛋白质的SEQ ID的“protein_id”限定符

特征键和限定符

核苷酸序列：“CDS”特征键

- “CDS”特征定位可以使用“join”定位操纵基因将不连续的序列片段连为一个编码区

`join(location1,location2)`

- “CDS”特征定位可以用“complement”操纵基因来表明该特征位于定位描述符所明确的序列的互补链上

`complement(location)`

特征键和限定符

氨基酸序列：常用特征键

特征键	描述	强制限定符?
SITE	表明有趣的单个氨基酸位点	必须用强制限定符NOTE来说明位点
REGION	表明感兴趣区	无；NOTE为可选内容
BINDING	表明化学基团的结合位点	强制限定符NOTE中必须包含化学基团名称
UNSURE	表明序列中的不确定区	无；NOTE为可选内容

特征键和限定符

氨基酸序列：改性氨基酸

- “改性氨基酸”是除以下外的任何氨基酸：

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, 第3段(e))

- 在可能的情况下，“改性核苷酸”应以对应的未改性核苷酸表示。否则，可以以“X”来表示。例如，在序列中，“羟赖氨酸”应以“K”来表示。“鸟氨酸”应以“X”来表示。(ST.26, 第29段)
- 符号“n”仅等于一个残基。

特征键和限定符

氨基酸序列：改性氨基酸

- 可以使用若干特征键来表示“改性氨基酸”：

特征键	描述
SITE	表示不是翻译后的改性氨基酸
MOD_RES	表示是翻译后的改性氨基酸
CARBOHYD	表示是糖基化氨基酸
LIPID	表示脂质部分与氨基酸的共价结合

- 以上每个特征键都必须包含值为修饰描述的NOTE强制限定符。

特征定位格式

定位格式

所有分子类型的定位描述符

- 定位描述符用于确定序列中特征的位置
- ST.26规定了定位描述符格式的强制性要求
- 下列定位描述符格式可用于核苷酸和氨基酸序列：

定位描述符类型	语法	说明
单个残基编号	x	指序列中的单个残基。
界定序列范围的残基号	x..y	指包含起始残基和末端残基并由其所约束的连续残基范围。
位于第一个指定残基号之前或最后一个指定残基号之后的残基	<x >x <x..y x..>y <x..>y	指包含特定残基或残基段并且范围超过特定残基的区。'<'和'>'符号可以用于单个残基或残基段的起始和末端残基号，来表明某个特征的范围超过了特定残基号。

定位格式

所有分子类型的定位描述符

定位样例	说明
467	指序列中的残基467。
340..565	指包含残基340和残基565并由其所约束的连续残基范围。
<1	指第一个残基之前的特征定位。
<345..500	表明特征的确切下限未知。定位始自345之前的某个残基，持续至并且包括残基500。
<1..888	表明特征始自序列中的第一个残基之前，持续至并且包括残基888。
1..>888	表明特征始自序列中的第一个残基，持续至残基888之后。
<1..>888	表明特征始自序列中的第一个残基之前，持续至残基888之后。

WIPO标准ST.26, 第70段 (a)

定位格式

仅用于核苷酸序列的定位描述符

- 以下定位描述符格式仅可用于DNA和RNA序列：

定位描述符类型	语法	说明
两个相邻核苷酸之间的位点	x^y	指两个相邻核苷酸之间的位点，例如内切核苷酸的切割位点。相邻核苷酸的位置号以 \wedge 隔开。该描述符的格式可以是 x^x+1 （例如55 \wedge 56），或对环形核苷酸而言， x^1 ，其中“x”是分子总长度，例如1000 \wedge 1表示长度为1000的环形核苷酸。

WIPO标准ST.26, 第66段 (b)

定位格式

仅用于核苷酸序列的定位描述符

- 使用 x^y 定位格式时， x 和 y 必须是相邻残基。

定位样例	说明
123 ¹ 124	指残基123和124之间的位点。
867 ¹	在带有残基867的环形分子中，指标示为位置1的残基和标示为位置的残基之间的位点。

WIPO标准ST.26, 第70段 (b)

定位格式

定位描述符——氨基酸序列特例

- 用于“CROSSLINK”或“DISULFID”特征键时，定位描述符x..y表明两个标示出的残基之间的链内链接

定位描述符类型	语法	说明
通过链内交叉链接而连接的残基号	x..y	用于“CROSSLINK”或“DISULFID”等表示链内交叉链接的特征时，指通过链内链接连接的氨基酸。

WIPO标准ST.26, 第66段 (c)

定位格式

定位描述符——氨基酸序列特例

```

<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_qual>
          <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_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>DISULFID</INSDFeature_key>
        <INSDFeature_location>30..50</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier id="q22">
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>disulfide bond</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>QKKMIQFFKITHRYYYDIIIEHLCAKYDMNSVISNALFAKLNLMQYTDGLSTHEKIIINTSNPLTFSIVISLQRCVINLGGST</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>

```

定位格式

核苷酸序列的定位运算符

- DNA和RNA序列可以使用三个定位运算符：“join”、“order”和“complement”
- 使用合并定位运算符意味着，由定位描述符进行说明的多个核苷酸残基通过生物过程进行了物理接触（ST.26，第68段）
- 使用“join”和“order”的定位必须有至少两个由逗号分隔的定位描述符
- “complement”可以与“join”或“order”共同使用

定位语法	定位说明
join(location,location,...,location)	标示出的位置被连接起来（以端到端的方式放置），形成一个连续的序列。
order(location,location,...,location)	以指明的顺序找到各种元素，但对连接这些元素是否合理不做任何暗示。
complement(location)	表示当按5'至3'的方向或按模仿5'至3'的方向读取时，该特征位于与位置描述符指明的序列段互补的链上。

定位格式

核苷酸序列的定位运算符

定位样例	说明
<code>join(12..78,134..202)</code>	表明12至78区和134至202区应被连接起来，形成一个连续的序列。
<code>order(15..228,341..502)</code>	表明15至228区和341至502区以所示顺序表示
<code>complement(34..126)</code>	由核苷酸126的互补核苷酸起始，在核苷酸34的互补核苷酸终止（特征位于所示链的互补链上）。
<code>complement(join(2691..4571, 4918..5163))</code>	连接核苷酸2691至4571和4918至5163，然后与所连接的片段形成互补（特征位于所示链的互补链上）。
<code>join(complement(4918..5163), complement(2691..4571))</code>	与4918至5163和2691至4571区形成互补，然后将互补的片段连接起来（特征位于所示链的互补链上）。

WIPO标准ST.26, 第70段 (b)

提问环节

限定符的值

限定符的值

- 限定符进一步界定特征。

限定符的值

- 限定符进一步界定特征。
- 限定符包含限定符名称，通常还包含限定符的值。

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

限定符的值

- 限定符进一步界定特征。
- 限定符包含限定符名称，通常还包含限定符的值。

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

- 每个特征键都会有允许用于该特征的一系列限定符。有些特征键具有强制限定符。

限定符的值

- 限定符进一步界定特征。
- 限定符包含限定符名称，通常还包含限定符的值。

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

- 每个特征键都会有允许用于该特征的一系列限定符。有些特征键具有强制限定符。
- 特征键及其允许使用的限定符列于ST.26附件一第5部分（核苷酸序列）以及第7部分（氨基酸序列）。

限定符的值

- 例如——“misc_binding”特征键有一个强制限定符“bound_moiety”，以及六个可选限定符：

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附件一第5.12部分)

限定符的值

- 限定符进一步界定特征。
- 限定符包含限定符名称，通常还包含限定符的值。

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

- 每个特征键都会有允许用于该特征的一系列限定符。有些特征键具有强制限定符。
- 特征键及其允许使用的限定符列于ST.26附件一第5部分（核苷酸序列）以及第7部分（氨基酸序列）。
- 核苷酸序列的限定符及其说明列于ST.26附件一第6部分（核苷酸序列）以及第8部分（氨基酸序列）。

限定符的值

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附件一第6.3和6.22部分)

限定符的值

格式类型

■ 限定符的值有以下几种格式类型：

1. 具有可选预定义值的限定符；
2. 具有已定义值格式的限定符；
3. 值为序列的限定符；
4. 值为**NO**的限定符；
5. 值为“自由文本”的限定符

——“自由文本”限定符值的一个子集被归类为“语种相关”

限定符的值

格式类型——可选预定义值

- 具有可选预定义值的限定符
- 例如：

“codon_start”——值可以为“1”、“2”或“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>

限定符值

格式类型——可选预定义值

■ 例如：“rpt_type”具有有限的可选值集

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;

限定符的值

格式类型——可选预定义值

- 对于具有可选预定义值的限定符，WIPO Sequence将在预填充的下拉菜单中提供所有被允许的值：

The screenshot displays the 'QUALIFIERS' section of the WIPO Sequence interface. A dropdown menu is open for the 'rpt_type' qualifier, showing a list of pre-defined values. The 'dispersed' value is currently selected and highlighted in blue. Below the dropdown, the 'SEQUENCE' section shows a grid of nucleotide sequences.

QUALIFIERS

Add qualifier

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

SEQUENCE

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

180

限定符的值

格式类型——已定义值格式

- 具有已定义值格式的限定符

- 例如：“anticodon”——值必须是以下格式：

“(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	<pre><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></pre>

限定符的值

格式类型——序列

- 值为序列的限定符
- 例如：“translation”——值必须为使用单字母氨基酸缩写的序列

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.

限定符的值

格式类型——序列

- 限定符“replace”——值可以是单个核苷酸残基、残基序列或空
- “replace”空值表示删除了对应特征中所标明的残基

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	<pre><INSDQualifier_value>a</INSDQualifier_value> <INSDQualifier_value></INSDQualifier_value> - for a deletion</pre>

限定符的值

格式类型——序列

- 限定符“replace”——值可以是单个核苷酸残基、残基序列或空
- “replace”空值表示删除了对应特征中所标明的残基

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	<pre><INSDQualifier value>a</INSDQualifier value> <INSDQualifier_value></INSDQualifier_value> - for a deletion</pre>

“空值”

限定符的值

格式类型——NO值

- 值为NO的限定符

- 例如：

“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将不会允许为空值限定符添加值
- 这些限定符不能有空值元素 “INSDQualifier_value”

限定符的值

格式类型——“自由文本”

- 许多限定符值格式为“自由文本”
- **ST.26第3段（n）**将“自由文本”定义为“……特定限定符的一种值格式，表现形式为描述性的文本短语或其他特定格式（如附件一所示）”
- 自由文本限定符值的上限是**1000**个字符（**ST.26，第86段**）
- 值格式为“自由文本”的限定符子集为“**语种相关**”
- “语种相关”自由文本“可能需要翻译国家、地区或国际程序”（**ST.26，第3段（o）**）

限定符的值

格式类型——“自由文本”

- 值为“语种相关自由文本”的限定符可能需要翻译
- 如何确定具有自由文本值格式的限定符是否为语种相关？
 - **ST.26**附件一第**6**部分表**5**列出了所有具有语种相关自由文本值格式的核苷酸序列限定符
 - **ST.26**附件一第**8**部分表**6**列出了所有具有语种相关自由文本值格式的氨基酸序列限定符
 - 见限定符说明中的“强制性值格式”

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>

限定符的值

格式类型——“自由文本”

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	<pre><INSDQualifier_value>23/108</INSDQualifier_value> <INSDQualifier_value>1 in 12</INSDQualifier_value> <INSDQualifier_value>0.85</INSDQualifier_value></pre>
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	<pre><INSDQualifier_value>essential for recognition of cofactor </INSDQualifier_value></pre>
Comment	The function qualifier is used when the gene name and/or product name do not convey the function attributable to a sequence.

限定符的值

格式类型——“语种相关自由文本”

- “语种相关自由文本”限定符值可以以序列表XML中的两种语言提供：英语和另一种非英语语言（**ST.26**，第87段）

限定符的值

格式类型——“语种相关自由文本”

- “语种相关自由文本”限定符值可以以序列表XML中的两种语言提供：英语和另一种非英语语言（**ST.26**，第87段）
- 语种相关的限定符值如果是英语的，就必须在INSDQualifier_value元素中提供

限定符值

格式类型——“语种相关自由文本”

- “语种相关自由文本”限定符值可以以序列表XML中的两种语言提供：英语和另一种非英语语言（**ST.26**，第87段）
- 语种相关的限定符值如果是英语的，就必须在**INSDQualifier_value**元素中提供
- 语种相关的限定符值如果是英语以外的任何其他语言，则必须在**NonEnglishQualifier_value**元素中提供

限定符值

格式类型——“语种相关自由文本”

- “语种相关自由文本”限定符值可以以序列表XML中的两种语言提供：英语和另一种非英语语言（**ST.26**，第87段）
- 语种相关的限定符值如果是英语的，就必须在**INSDQualifier_value**元素中提供
- 语种相关的限定符值如果是英语以外的任何其他语言，则必须在**NonEnglishQualifier_value**元素中提供
- 仅允许具有语种相关自由文本值格式的限定符使用**NonEnglishQualifier_value**元素（**ST.26**，第87段（b））

限定符值

格式类型——“语种相关自由文本”

```

<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_qual>
      - <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_qual>
  </INSDFeature>
</INSDSeq_feature-table>
  <INSDSeq_sequence>GSLSDVRKDVEKRIDKALEAFKNKMDKEK</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="3">

```

限定符值

格式类型——“语种相关自由文本”

■ 当序列列表XML中包含非英语限定符值时：

1. 序列列表XML的根元素必须包含具有适当双字母语言代码缩写的“nonEnglishFreeTextLanguageCode”属性（ST.26，第43段和第87段（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">
```

限定符值

格式类型——“语种相关自由文本”

■ 当序列XML中包含非英语限定符值时：

1. 序列XML的根元素必须包含具有适当双字母语言代码缩写的“nonEnglishFreeTextLanguageCode”属性（ST.26，第43段和第87段（b））；
2. 序列中的所有语种相关限定符的值都必须使用“nonEnglishFreeTextLanguageCode”属性中所表明语言；

限定符值

格式类型——“语种相关自由文本”

■ 当序列XML中包含非英语限定符值时：

1. 序列XML的根元素必须包含具有适当双字母语言代码缩写的“nonEnglishFreeTextLanguageCode”属性（ST.26，第43段和第87段（b））；
2. 序列中的所有语种相关限定符的值都必须使用“nonEnglishFreeTextLanguageCode”属性中所表明的语言；
3. 当一个限定符既有NonEnglishQualifier_value也有INSDQualifier_value时，两个元素中所含的信息必须相同（ST.26，第87段（c））。

限定符值

格式类型——“语种相关自由文本”

- INSDQualifier的“id”属性是什么？
- ST.26第87段（d）称，“对于语种相关限定符，INSDQualifier元素可包含可选属性id。该属性的值格式必须为“q”后接正整数，例如“q23”，并且是该INSDQualifier元素所独有的，例如该属性值在一份序列列表文件中仅可使用一次。”

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

限定符值

格式类型——“语种相关自由文本”

■ INSDQualifier “id”属性:

- 单独明确可能需要翻译的限定符值，以由WIPO Sequence导出XLIFF文件。
- 可选;
- 仅被允许用于语种相关限定符;
- 在序列表中必须是独一无二的;
- 自动添加至由WIPO Sequence生成的XML;

限定符值

格式类型——“语种相关自由文本”

■ “originalFreeTextLanguageCode”属性

```
<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第43段将“originalFreeTextLanguageCode”属性定义为“……语种相关自由文本限定符所用的单一原始语言的语言代码”

■ 该属性为可选属性

特殊情况：
DNA/RNA杂合分子
DNA中的尿嘧啶和
RNA中的胸腺嘧啶

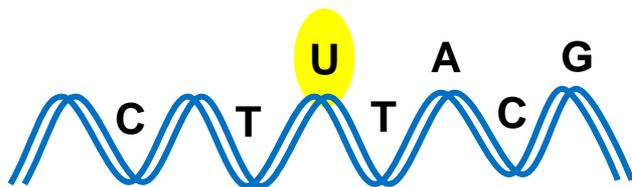
特殊情况

- 提醒：ST.26不允许使用符号“u”来代表尿嘧啶
- 在DNA中，“t”是胸腺嘧啶
- 在RNA中，“t”是尿嘧啶
- 需考虑两种情况：
 1. 含有尿嘧啶核碱基的DNA分子，或含有胸腺嘧啶核碱基的RNA分子；
 2. DNA/RNA杂合分子

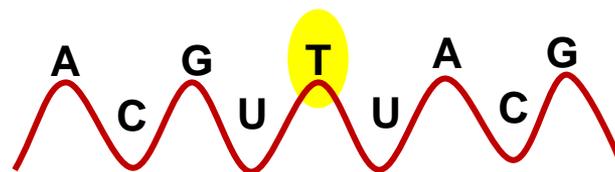
特殊情况

DNA中的尿嘧啶和RNA中的胸腺嘧啶

- 如果序列中存在含有尿嘧啶核碱基的DNA骨架，或存在含有胸腺嘧啶核碱基的RNA骨架——则说明为“改性核苷酸”。



DNA



RNA

适用ST.26第14段：

“14. 符号“t”将在DNA中被翻译为胸腺嘧啶，在RNA中被翻译为尿嘧啶。DNA中的尿嘧啶或RNA中的胸腺嘧啶被视为**改性核苷酸**，必须使用第19段中提供的特征表进一步说明。”

特殊情况

DNA和RNA序列

一件申请公开了以下RNA序列：

5' -cgucccacgugtccgaggua -3'

- 注意位置12的“胸腺嘧啶”。该残基必须被注释为改性核苷酸。
- ST.26第19段称：DNA中的尿嘧啶或RNA中的胸腺嘧啶被视为改性核苷酸，在序列中必须用“t”来表示，并且必须在特征表中使用特征键“modified_base”和值为“OTHER”的限定符“mod_base”以及值分别为“uracil”或“thymine”的限定符“note”进一步说明。

特殊情况

DNA和RNA序列

一件申请公开了以下RNA序列：

5' -cgucccacgugtccgaggua -3'

- 注意位置12的“胸腺嘧啶”。该残基必须被注释为改性核苷酸。
- ST.26第19段称：DNA中的尿嘧啶或RNA中的胸腺嘧啶被视为改性核苷酸，在序列中必须用“t”来表示，并且必须在特征表中使用特征键“**modified_base**”和值为“OTHER”的限定符“**mod_base**”以及值分别为“uracil”或“thymine”的限定符“**note**”进一步说明。

特殊情况

DNA和RNA序列

一件申请公开了以下RNA序列：

5' -cgucccacgugtccgaggua-3'

- ✓ 所有尿嘧啶残基都必须用符号“t”来表示。因此，该序列必须在序列表中表示为：

cgtcccacgtgtccgaggta

特殊情况

DNA和RNA序列

一件申请公开了以下RNA序列：

5' -cgucccacgugtccgaggua-3'

- ✓ 所有尿嘧啶残基都必须用符号“t”来表示。因此，该序列必须在序列表中表示为：

cgtcccacgtgtccgaggta

- ✓ 位置为“12”的特征键“modified_base”

特殊情况

DNA和RNA序列

一件申请公开了以下RNA序列：

5' -cgucccacgugtccgaggua-3'

- ✓ 所有尿嘧啶残基都必须用符号“t”来表示。因此，该序列必须在序列表中表示为：

cgtccccacgtgtccgaggta

- ✓ 位置为“12”的特征键“modified_base”
- ✓ 值为“OTHER”的限定符“mod_base”

特殊情况

DNA和RNA序列

一件申请公开了以下RNA序列：

5' -cgucccacgugtccgaggua-3'

- ✓ 所有尿嘧啶残基都必须用符号“t”来表示。因此，该序列必须在序列表中表示为：

cgtccccacgtgtccgaggta

- ✓ 位置为“12”的特征键“modified_base”
- ✓ 值为“OTHER”的限定符“mod_base”
- ✓ 值为“thymine”的限定符“note”

特殊情况

DNA和RNA序列

```

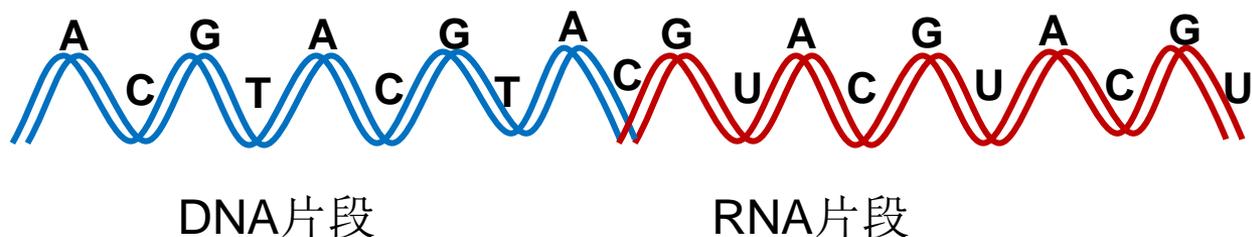
<SequenceData sequenceIDNumber="3">
  <INSDSeq>
    <INSDSeq_length>20</INSDSeq_length>
    <INSDSeq_moltype>RNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..20</INSDFeature_location>
        <INSDFeature_qual>
          <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_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>modified_base</INSDFeature_key>
        <INSDFeature_location>12</INSDFeature_location>
        <INSDFeature_qual>
          <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_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>cgtcccacgtgtcgcgagta</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>

```

特殊情况

DNA/RNA杂合序列

- 如果序列为杂合分子；例如骨架一部分为DNA，一部分为RNA：



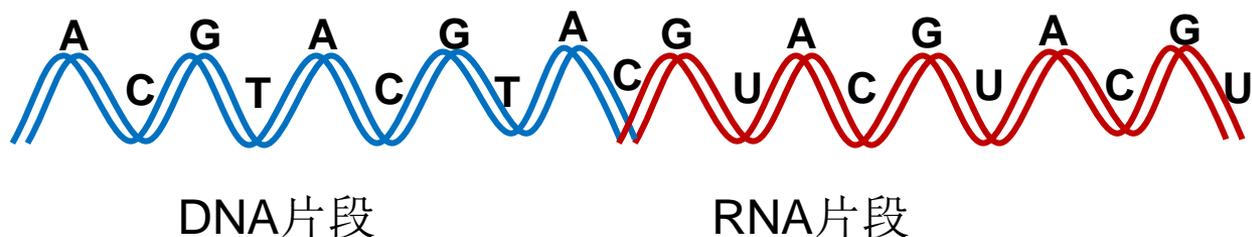
适用ST.26第55段：

“55. 对于同时包含一个或多个核苷酸的DNA和RNA片段的核苷酸序列，分子类型必须表明为DNA。DNA/RNA组合分子必须在特征表中使用特征键“source”和值为“synthetic construct”的强制限定符“organism”以及值为“other DNA”的强制限定符“mol_type”进一步说明。DNA/RNA组合分子的每个DNA和RNA片段必须用特征键“misc_feature”和“note”限定符进一步说明，明确该片段是DNA还是RNA。”

特殊情况

DNA/RNA杂合序列

- 如果序列为杂合分子；例如骨架一部分为DNA，一部分为RNA：



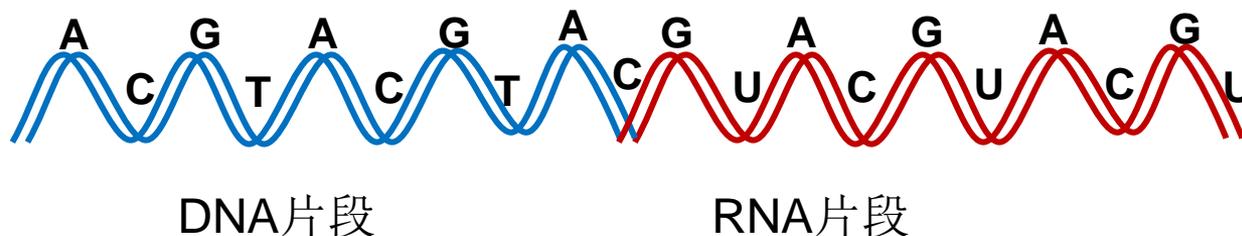
适用ST.26第55段：

“55. 对于同时包含一个或多个核苷酸的DNA和RNA片段的核苷酸序列，**分子类型必须表明为DNA**。DNA/RNA组合分子必须在特征表中使用特征键“source”和值为“synthetic construct”的强制限定符“organism”以及值为“other DNA”的强制限定符“mol_type”进一步说明。DNA/RNA组合分子的每个DNA和RNA片段必须用特征键“misc_feature”和“note”限定符进一步说明，明确该片段是DNA还是RNA。”

特殊情况

DNA/RNA杂合序列

- 如果序列为杂合分子；例如骨架一部分为DNA，一部分为RNA：



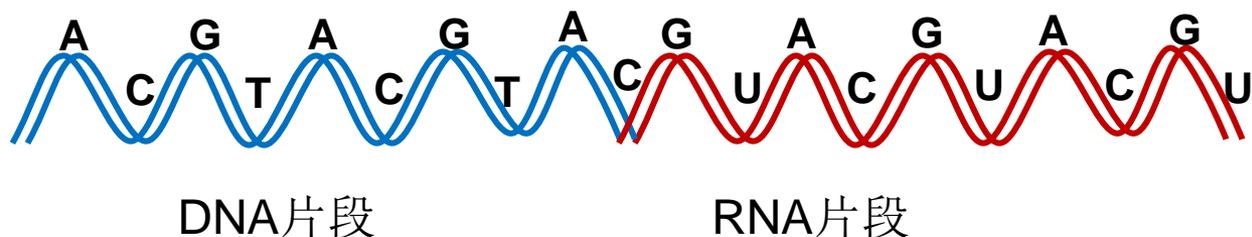
适用ST.26第55段：

“55. 对于同时包含一个或多个核苷酸的DNA和RNA片段的核苷酸序列，分子类型必须表明为DNA。DNA/RNA组合分子必须在特征表中使用特征键“source”和值为“synthetic construct”的强制限定符“organism”以及值为“other DNA”的强制限定符“mol_type”进一步说明。DNA/RNA组合分子的每个DNA和RNA片段必须用特征键“misc_feature”和“note”限定符进一步说明，明确该片段是DNA还是RNA。”

特殊情况

DNA/RNA杂合序列

- 如果序列为杂合分子；例如骨架一部分为DNA，一部分为RNA：



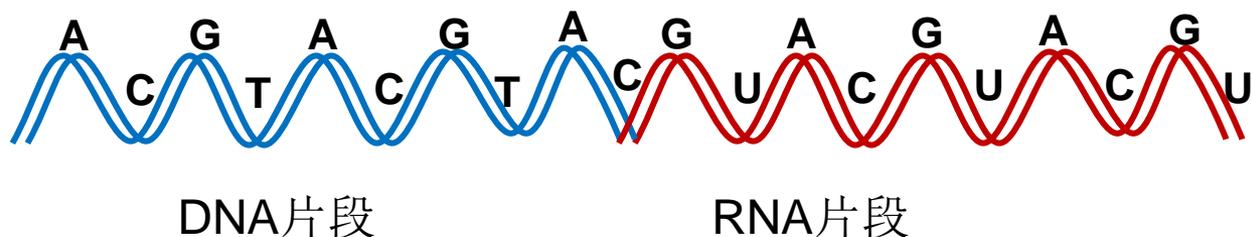
适用ST.26第55段：

“55. 对于同时包含一个或多个核苷酸的DNA和RNA片段的核苷酸序列，分子类型必须表明为DNA。DNA/RNA组合分子必须在特征表中 使用特征键“source”和值为“synthetic construct”的强制限定符“organism”以及值为“other DNA”的强制限定符“mol_type”进一步说明。DNA/RNA组合分子的每个DNA和RNA片段必须用特征键“misc_feature”和“note”限定符进一步说明，明确该片段是DNA还是RNA。”

特殊情况

DNA/RNA杂合序列

- 如果序列为杂合分子；例如骨架一部分为DNA，一部分为RNA：



适用ST.26第55段：

“55. 对于同时包含一个或多个核苷酸的DNA和RNA片段的核苷酸序列，分子类型必须表明为DNA。DNA/RNA组合分子必须在特征表中使用特征键“source”和值为“synthetic construct”的强制限定符“organism”以及值为“other DNA”的强制限定符“mol_type”进一步说明。DNA/RNA组合分子的每个DNA和RNA片段必须用特征键“misc_feature”和“note”限定符进一步说明，明确该片段是DNA还是RNA。”

特殊情况

DNA/RNA杂合序列

一件申请公开了以下DNA/RNA杂合序列：

5' -ACCTGCcgucccacguguccgagguaGCATTA-3'

其中大写符号表示DNA部分，小写符号表示RNA部分。

■ 残基1-6和27-32是DNA，残基7-26是RNA

■ 考虑：

1. 生物体名称
2. 分子类型和mol_type
3. DNA和RNA片段识别

特殊情况

DNA和RNA序列

一件申请公开了以下DNA/RNA杂合序列：

5' -ACCTGCcgucccacguguccgagguaGCATTA-3'

其中大写符号表示DNA部分，小写符号表示RNA部分。

- ST.26第55段称：“.....分子类型必须表明为DNA。DNA/RNA组合分子必须在特征表中使用特征键“source”和值为“synthetic construct”的强制限定符“organism”以及值为“other DNA”的强制限定符“mol_type”进一步说明。”



分子类型= “DNA”



限定符organism name = “synthetic construct”



限定符mol_type = “other DNA”

特殊情况

DNA和RNA序列

一件申请公开了以下DNA/RNA杂合序列：

5' -ACCTGCcgucccacguguccgagguaGCATTA-3'

其中大写符号表示DNA部分，小写符号表示RNA部分。

- ✓ 所有尿嘧啶残基都必须用符号“t”来表示。因此，该序列必须在序列表中表示为：

acctgccgtcccacgtgtccgaggtagcatta

特殊情况

DNA和RNA序列

一件申请公开了以下DNA/RNA杂合序列：

5' -ACCTGCcgucccacguguccgagguaGCATTA-3'

其中大写符号表示DNA部分，小写符号表示RNA部分。

- ST.26第55段称：“DNA/RNA组合分子的每个DNA和RNA片段必须用特征键“misc_feature”和限定符“note”进一步说明，明确该片段是DNA还是RNA。”

✓ 三个片段 = 三个“misc_feature”特征键

特殊情况

DNA和RNA序列

一件申请公开了以下DNA/RNA杂合序列：

5' -ACCTGCcgucccacguguccgagguaGCATTA-3'

其中大写符号表示DNA部分，小写符号表示RNA部分。

- ✓ 片段1，残基1-6：
位置为“1..6”的“misc_feature”特征键
值为“DNA”的“note”限定符
- ✓ 片段2，残基7-26：
位置为“7..26”的“misc_feature”特征键
值为“RNA”的“note”限定符
- ✓ 片段3，残基27-32：
位置为“27..32”的“misc_feature”特征键
值为“DNA”的“note”限定符

特殊情况

DNA和RNA序列

```

<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>
    </INSDSeq_feature-table>
  </INSDSeq>
</SequenceData>

```


问答环节

核苷酸类似物、D-氨基酸 和分支序列

新分子类型

核苷酸类似物

- 包含一个或多个核苷酸类似物的核酸序列须符合ST.26的规定
- ST.26第3段（g）（2）“核苷酸”的定义中包含核苷酸类似物相关内容：

“2'-脱氧核糖-5'-单磷酸或核糖-5'-单磷酸的类似物在形成核酸类似物的骨架时，导致核碱基的排列模拟包含2'-脱氧核糖-5'-单磷酸或核糖-5'-单磷酸骨架的核酸的核碱基排列，在此情况下核酸类似物能够与互补核酸形成碱基对”

- 常见核酸类似物包括肽核酸（PNA）、甘油核酸（GNA）、苏糖核酸和吗啉基。
- 必须模拟5'至3'的方向从左至右进行表示（ST.26，第11段）

新分子类型

核苷酸类似物

一件专利申请公开了以下甘油核酸（GNA）序列：

PO_4 -tagttcattgactaaggctccccattgact-OH

其中序列的 PO_4 端模拟了一个DNA序列的5'端。

✓ 要求将该序列纳入序列表中

新分子类型

核苷酸类似物

一件专利申请公开了以下甘油核酸（GNA）序列：

PO_4 -tagttcattgactaaggctccccattgact-OH

其中序列的 PO_4 端模拟了一个DNA序列的5'端。

- ✓ 要求将该序列纳入序列表中
- ✓ PO_4 端模拟了5'端，因此该序列必须以所示方向表示

新分子类型

核苷酸类似物

一件专利申请公开了以下甘油核酸（GNA）序列：

PO_4 -tagttcattgactaaggctccccattgact-OH

其中序列的 PO_4 端模拟了一个DNA序列的5'端。

- ✓ 要求将该序列纳入序列表中
- ✓ PO_4 端模拟了5'端，因此该序列必须以所示方向表示
- ✓ 完整序列必须在注释中纳入以下内容：“modified_base”特征键、值为“OTHER”的“mod_base”限定符和包含改性核苷酸未经缩略的完整名称（例如“甘油核酸”）的note限定符。

新分子类型

核苷酸类似物

```

- <INSDFeature>
  <INSDFeature_key>modified_base</INSDFeature_key>
  <INSDFeature_location>1..30</INSDFeature_location>
  - <INSDFeature_qual>
    - <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_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>tagttcattgactaaggctccccattgact</INSDSeq_sequence>
</INSDSeq>

```

注：可在WIPO标准ST.26附件六例3 (g) -4中找到对此例的延伸讨论。

新分子类型

D-氨基酸

- 包含一个或多个D-氨基酸的氨基酸序列须符合ST.26的规定

- ST.26第3段（a）“氨基酸”的定义中包含D-氨基酸相关内容：

““氨基酸”指的是可以用附件一所载任何符号表示的任何氨基酸（见第3部分，表3）。这些氨基酸包括D-氨基酸和含有改性或合成侧链的氨基酸。”

- 在可能的情况下，D-氨基酸必须在序列中以对应的未改性L-氨基酸符号表示
- 必须在特征表中说明为改性氨基酸

新分子类型

D-氨基酸

一件专利申请对以下序列进行了说明：

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

✓ 要求将该序列纳入序列表中

新分子类型

D-氨基酸

一件专利申请对以下序列进行了说明：

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

- ✓ 要求将该序列纳入序列表中
- ✓ 应表示为：AGKLGGM
- ✓ 位置1的丙氨酸、位置2的谷氨酸和位置6的甲硫氨酸必须分别以“SITE”特征键和包含对应氨基酸未经缩略的完整名称的“NOTE”限定符进行注释

新分子类型

D-氨基酸

```

- <INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>1</INSDFeature_location>
  - <INSDFeature_qual>
    - <INSDQualifier id="q4">
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>D-alanine</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
- <INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>2</INSDFeature_location>
  - <INSDFeature_qual>
    - <INSDQualifier id="q5">
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>D-glutamic acid</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
- <INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>6</INSDFeature_location>
  - <INSDFeature_qual>
    - <INSDQualifier id="q6">
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>D-methionine</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>AEKLG</INSDSeq_sequence>

```

注：可在WIPO标准ST.26附件六例3 (a) -1中找到对类似例子的延伸讨论

新分子类型

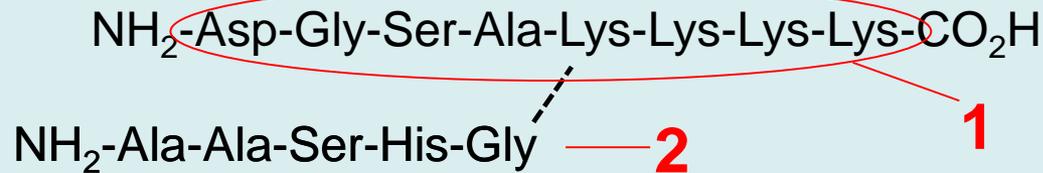
分支序列

- 分支核酸序列和分支氨基酸序列须符合**ST.26**的规定
- 如果分支序列的线性区包含**10个**或更多明确界定的核苷酸或**4个**或更多明确界定的氨基酸，则这些线性区必须被纳入序列表中（**WIPO标准ST.26**，第7段）
- 满足长度要求下限的分支序列各线性区必须作为单独序列纳入，并且有其单独的**SEQ ID**编号
- 所考虑的必须是每个单独线性区中明确界定的残基数量，而非整个结构中明确界定的残基总数

新分子类型

分支序列

一件专利申请用以下序列对一个肽进行了说明：



此处——表示在甘氨酸的碳端和赖氨酸的侧链之间的酰胺键

- ✓ 两个线性区都含有超过4个明确界定的氨基酸，因此它们都必须被纳入序列表中
- ✓ 每个线性区都必须作为单独序列纳入，并且拥有其单独的SEQ ID编号

新分子类型

分支序列

1

```

- <INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>5</INSDFeature_location>
  - <INSDFeature_qual>
    - <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_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>DGS AKKKK</INSDSeq_sequence>

```

2

```

- <INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>5</INSDFeature_location>
  - <INSDFeature_qual>
    - <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_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>AASHG</INSDSeq_sequence>

```

注：可在WIPO标准ST.26附件六例7 (b) -3中找到对类似例子的延伸讨论

序列变异

序列变异

- ST.26第3段（m）将“变异序列”定位为

“包含一个或多个与主序列的**差异**的核苷酸或氨基酸序列。这些差异可能包括不同残基（见第15和27段）、改性残基（见第3段（g）、第3段（h）、第16段和第29段）、删减、插入和替代。见第93至95段。”

- 公开变异序列的方式将决定其必须在序列表中如何表示。
- ST.26第93至95段规定了变异必须如何表示。

序列变异

第93段

第93段：以列举其残基的方式分别公布并且包含在第7段中的主序列和该序列的任何变异必须各自纳入序列表中，并为其分配其单独的序列识别编号。



如果各变异分别列举，则每个变异必须有其单独的SEQ ID编号！

序列变异

第93段

一件专利申请中包含具有以下多重序列比对的图：

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

小写字母表示比对序列中的主要氨基酸残基。

- ✓ 六个列举出的序列必须各自作为单独序列纳入序列表中，并且各自有单独的SEQ ID编号。

注：可在WIPO标准ST.26附件六例93-3中找到对类似例子的延伸讨论。

序列变异

适当注释

- 为变异作注释的适当特征键取决于分子类型和变异性质：

序列类型	特征键	限定符	使用
核苷酸	variation	replace or note	自然发生的突变和多态性，例如等位基因、限制性片段长度多态性。
核苷酸	misc_difference	replace or note	人工引入的变化，例如通过基因操纵或化学合成。
氨基酸	VAR_SEQ	NOTE	通过选择性剪接、选择性使用启动子、选择性起始和核糖体框架位移产生的变异。
氨基酸	VARIANT	NOTE	不适用VAR_SEQ的任何类型的变异。

WIPO标准ST.26, 第96段

序列变异

最具限制性的模糊符号

- ST.26第15段和第27段称，需要模糊符号时，“应使用最具限制性的符号……”

这是什么意思？

例如：

在一个核苷酸序列中，如果一个位置可以是“a”或“c”，使用模糊符号“m”，而非“n”。

在一个氨基酸序列中，如果一个位置可以是“L”或“I”，使用模糊符号“J”，而非“X”。

请记住“n”和“X”有默认值，因此如果“n”或“X”在任何时候被用于默认值之外的内容，需要对其作出注释。

序列变异

第94段

第94段：作为列举了一个或多个位置上不同残基的单独序列公开的任何变异序列必须纳入序列表中，并且应以单独序列来表示，在此情况下，列举的其他残基由最具限制性的模糊符号表示（见第15段和第27段）。



如果变异并没有分别列举，而只是在主序列中作为可变残基显示，就不需要有单独的SEQ ID编号！

序列变异

第94段

一件专利申请公开了序列中的一个肽：

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

- ✓ 可作为单独序列纳入序列表中
- ✓ 该序列的最佳表示为：GGGJATX
- ✓ [Leu or Ile] 应以最具限制性的模糊符号“J”来表示
- ✓ [Ser or Thr] 应以符号“X”来表示，辅以特征键“VARIANT”，并用限定符“note”说明X是丝氨酸还是苏氨酸。

注：可在WIPO标准ST.26附件六例94-1中找到对该例的延伸讨论。

序列变异

第95段

第95段：在序列列表的主序列中仅以提及删减、插入或替代进行公开的任何变异序列，应纳入序列列表中。纳入序列列表时，此类变异序列：

- (a) 如果包含单个位置或多个不同位置的变异，并且这些变异的出现是相互独立的，可由主序列的注释进行表示；
- (b) 如果包含多个不同位置的变异，并且这些变异的出现是相互独立的，应作为单独序列表示，并且被分配其单独的序列识别编码；以及
- (c) 如果包含含有超过1000个残基的插入序列或替代序列，必须作为单独序列表示，并且被分配其单独的序列识别编码（见第86段）。



“提及删减、插入或替代”意味着变异以白话形式公开。

序列变异

第95段 (a)

一件申请中包含以下公开：

肽片段1: Gly-Leu-Pro-Xaa-Arg-Ile-Cys

其中Xaa为任何氨基酸

* * *

.....在另一种呈现形式中，肽片段1为Gly-Leu-Pro-Xaa-Arg-Ile-Cys wherein Xaa can be Val, Thr, or Asp.....

* * *

.....在又一种呈现形式中，肽片段1为Gly-Leu-Pro-Xaa-Arg-Ile-Cys wherein Xaa can be Val...

主序列Gly-Leu-Pro-Xaa-Arg-Ile-Cys包含单个位置上的变异，并且这些变异的出现是相互独立的。

序列变异

第95段 (a)

- ✓ 主序列必须包含在序列表中，变异序列可作为主序列的注释进行表示
- ✓ 最具包容性的表现形式必须是序列表中所包含的版本——在此例中，即Xaa为“任何氨基酸”的版本
- ✓ 该序列必须表示为GLPXRIC，并且需要在位置4的特征键“VARIANT”以及说明Xaa为“任何氨基酸”的限定符“note”
- ✓ 尽管不作要求，但建议将下列三种变异作为单独序列纳入序列表中：

GLPVRIC

GLPTRIC

GLPDRIC

注：可在WIPO标准ST.26附件六例95 (a) -1中找到对该例的延伸讨论。

序列变异

第95段 (b)

一件专利申请说明了以下共有序列：

aatg n_1 cccacgaatg n_2 cac

其中 n_1 和 n_2 可以是a, t, g, or c

公开了以下若干变异序列：

如果 n_1 是a，则 n_2 是t、g或c

如果 n_1 是t，则 n_2 是a、g或c

如果 n_1 是g，则 n_2 是t、a或c

如果 n_1 是c，则 n_2 是t、g或a

共有序列含有多个不同位置的变异，并且这些变异的出现是相互独立的

序列变异

第95段 (b)

- ✓ 共有序列必须包含在序列表中，变异序列应作为单独序列进行表示
- ✓ 最具包容性的表现形式必须是序列表中所包含的版本：在此情况下， n_1 和 n_2 “可以是a、t、g或c”
- ✓ 该序列必须表示为：**aatgncccacgaatgncac**
- ✓ 无需为n作注释，因为在没有特征键的情况下，“n”被解释为“a”、“c”、“g”或“t”中的一个（见ST.26第15段）
- ✓ 尽管不作要求，但建议将下列四种变异作为单独序列纳入序列表中：

aatgacccacgaatgbcac (b = t, g, or c)

aatgtccccacgaatgvcac (v = a, g, or c)

aatggccccacgaatghcac (h = t, a, or c)

aatgccccacgaatgdcac (d = t, g, or a)

注：可在WIPO标准ST.26附件六例95 (b) 中找到对类似例子的延伸讨论。

序列变异

第95段 (c)

一件申请中公开了以下内容：

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

其中Xaa是甘氨酸或序列中插入的内容

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

一个包含含有超过1000个残基的插入序列或替代序列的变异序列

序列变异

第95段 (c)

- ✓ 其中Xaa = 插入超过1000个残基的变异序列必须作为具有其单独SEQ ID编号的单独序列纳入序列表中。
- ✓ 其中Xaa = 甘氨酸的变异序列也将作为具有其单独SEQ ID编号的单独序列纳入序列表中。

...-MGLPRGRICK-...

序列变异

Q: 哪一段适用于以下公开:

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

序列变异

Q: 哪一段适用于以下公开:

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

A: 第94段:

“作为列举了一个或多个位置上不同残基的单独序列公开的任何变异序列必须纳入序列表中，并且应以单独序列来表示，在此情况下，列举的不同残基由最具限制性的模糊符号表示。”

序列变异

Q: 哪一段适用于以下公开:

位置	1	2	3	4	5	6	7	8	9
序列	A	V	L	T	Y	L	R	G	E
变异1									A
变异2			P			P			
变异3			A	I	G	Y			
变异4							-		

表中的空格表示该变异中的氨基酸与“序列”中的对应氨基酸相同，“-”表示“序列”中的对应氨基酸已被删减。

序列变异

Q: 哪一段适用于以下公开:

位置	1	2	3	4	5	6	7	8	9
序列	A	V	L	T	Y	L	R	G	E
变异1									A
变异2			P			P			
变异3			A	I	G	Y			
变异4							-		

表中的空格表示该变异中的氨基酸与“序列”中的对应氨基酸相同，“-”表示“序列”中的对应氨基酸已被删减。

A: 第93段

“以列举其残基的方式分别公布并且包含在第7段中的主序列和该序列的任何变异必须各自纳入序列表中，并为其分配其单独的序列识别编号。”

序列变异

Q: 哪一段适用于以下公开:

图中公开了以下序列:

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

具体说明中包含关于序列1的以下信息:

.....在某些表现形式中, 序列1位置7的缬氨酸被丙氨酸替代.....

序列变异

Q: 哪一段适用于以下公开:

图中公开了以下序列:

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

具体说明中包含关于序列1的以下信息:

.....在某些表现形式中, 序列1位置7的缬氨酸被丙氨酸替代.....

A: 第95段 (a): “在序列列表的主序列中仅以提及删减、插入或替代进行公开的任何变异序列, 应纳入序列列表中。纳入序列列表时, 此类变异序列:

(a) 如果包含单个位置或多个不同位置的变异, 并且这些变异的出现是相互独立的, 可由主序列的注释进行表示;”

序列变异

Q: 哪一段适用于以下公开:

图中公开了以下序列:

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

具体说明中包含关于序列1的以下信息:

.....如果序列1位置7的缬氨酸被丙氨酸替代, 之后位置10的丙氨酸被缬氨酸替代.....

序列变异

Q: 哪一段适用于以下公开:

图中公开了以下序列:

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

具体说明中包含关于序列1的以下信息:

.....如果序列1位置7的缬氨酸被丙氨酸替代, 之后位置10的丙氨酸被缬氨酸替代.....

A: : 第95段 (a): “在序列列表的主序列中仅以提及删减、插入或替代进行公开的任何变异序列, 应纳入序列列表中。纳入序列列表时, 此类变异序列:

(b) 如果包含多个不同位置的变异, 并且这些变异的出现是相互独立的, 应作为单独序列表示, 并且被分配其单独的序列识别编码; ”

问答环节