

СТАНДАРТ ST. 26

Рекомендуемый Стандарт представления перечней нуклеотидных и аминокислотных последовательностей с использованием языка XML (Расширяемого языка разметки)

Версия 1.1

*Одобен Комитетом по Стандартам ВОИС (КСВ)
на своей пятой сессии 2 июня 2017*

Редакционное примечание подготовлено Международным бюро

На своей пятой сессии Комитет по Стандартам ВОИС (КСВ) согласился с тем, что переход от Стандарта ВОИС ST.25 к Стандарту ST.26 состоится в январе 2022 года. До того момента следует продолжать использовать Стандарт ST.25.

Стандарт публикуется с целью информирования Ведомств по промышленной собственности и других заинтересованных сторон.

Содержание

ВВЕДЕНИЕ.....	3
ОПРЕДЕЛЕНИЯ.....	3
ОБЛАСТЬ ПРИМЕНЕНИЯ.....	5
ОТСЫЛКИ.....	6
ПРЕДСТАВЛЕНИЕ ПОСЛЕДОВАТЕЛЬНОСТЕЙ.....	7
Нуклеотидные последовательности.....	7
Аминокислотные последовательности.....	11
Представление особых ситуаций.....	13
СТРУКТУРА ПЕРЕЧНЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ В XML.....	14
Корневой (основной) элемент.....	15
Часть общей информации.....	16
Часть с данными последовательности.....	20
Таблица характеристик.....	23
Функциональные ключи.....	23
Обязательные функциональные ключи.....	23
Местоположение признака.....	24
Квалификаторы признака.....	27
Обязательные квалификаторы признаков.....	27
Элементы квалификатора.....	27
Свободный текст.....	29
Кодирующие последовательности.....	30
Варианты.....	31

Приложения

- Приложение 1 – Контролируемая лексика
Приложение 2 – Определение типа документа для перечня последовательностей (DTD)
Приложение 3 – Образец перечня последовательностей (XML файл)
Приложение 4 – Подмножество символов из Таблицы Unicode Basic Latin Code
Приложение 5 – Дополнительные требования для обмена данными (только для патентных ведомств)
Приложение 6 – Руководство пользователя и Дополнение – Руководство по последовательностям в XML размещено по адресу:
<http://www.wipo.int/export/sites/www/standards/en/pdf/03-26-01.pdf> на страницах с 95 по 195.

СТАНДАРТ ST. 26

Рекомендуемый Стандарт представления перечней нуклеотидных и аминокислотных последовательностей с использованием языка XML (Расширяемого языка разметки)

Версия 1.1

*Одобен Комитетом по Стандартам ВОИС (КСВ)
на своей пятой сессии 2 июня 2017*

Введение

1. Данный Стандарт определяет требования к раскрытию в заявках на патент последовательностей нуклеотидов и аминокислот для включения в перечень последовательностей, представляет методы раскрытия и определение типа документа (Document Type Definition, далее - DTD) для перечней последовательностей в XML (eXtensible Markup Language (Расширяемого Языка Разметки)). Рекомендуется, чтобы ведомства по промышленной собственности принимали любые перечни последовательностей, соответствующие данному Стандарту, поданные как часть заявки на патент или касающиеся заявки на патент.

2. Цель данного Стандарта:

а) позволить заявителю единообразно составлять перечень последовательностей в заявке на патент, который был бы приемлем как для международных, так и для национальных или региональных процедур;

б) повысить достоверность и качество представления последовательностей для упрощения их распространения среди заявителей, широкой публики и экспертов,

с) облегчить поиск данных о последовательностях; и

д) обеспечить обмен данными о последовательностях в электронной форме, а также ввод соответствующих данных в компьютеризованные базы данных.

Определения

3. Для целей данного Стандарта выражение:

а) «аминокислота» означает любую аминокислоту, которая может быть представлена с помощью любого символа, определённого в Приложении 1 (см. раздел 3, Таблица 3). Такие аминокислоты включают, в частности, D-аминокислоты и аминокислоты, содержащие модифицированные или синтетические боковые цепи. Аминокислоты будут истолкованы как немодифицированные L-аминокислоты, если они далее не описаны в таблице характеристик как модифицированные в соответствии с параграфом 30. Для целей настоящего Стандарта остаток пептидной нуклеиновой кислоты (PNA) не считается аминокислотой, но считается нуклеотидом, как указано в параграфе 3 (g) (i) (2).

б) «контролируемая лексика» - это терминология, содержащаяся в Стандарте, которая

должна быть использована для описания свойств последовательностей, т.е. аннотирования области или интересующей позиции, определяемых в Приложении 1.

с) «перечисление остатков» означает раскрытие последовательности в патентной заявке путем перечисления каждого остатка последовательности, в которой:

- (i) остаток представляет собой имя, аббревиатуру, символ или структуру (например, HHHHHHQ или HisHisHisHisHisHisGln); или
- (ii) множество остатков представлено сокращенной формулой (например, His6Gln).

д) "намеренно пропущенная последовательность", также известная как пустая последовательность, обозначает метку-заполнитель (placeholder) сохранения нумерации последовательностей в перечне последовательностей для согласованности с раскрытием заявки, например, если последовательность удаляется из описания после подготовки перечня последовательностей.

е) «модифицированная аминокислота» означает любую аминокислоту, как описано в параграфе 3 (а), отличную от L-аланина, L-аргинина, L-аспарагина, L-аспарагиновой кислоты, L-цистеина, L-глутамина, L-глутаминовой кислоты, L-глицина, L-гистидина, L-изолейцина, L-лейцина, L-лизина, L-метионина, L-фенилаланина, L-пролина, L-пирролизина, L-серина, L-селеноцистеина, L-треонина, L-триптофана, L-тирозина или L-валина.

г) «модифицированный нуклеотид» означает любой нуклеотид, как описано в параграфе 3 (г), отличный от 3'-монофосфата дезоксиаденозина, 3'-монофосфата дезоксигуазидина, 3'-монофосфата дезоксцитидина, 3'-монофосфата дезокситимидина, 3'-монофосфата аденозина, 3'-монофосфата гуанозина, 3'- монофосфата цитидина или 3'-монофосфата уридина.

г) «нуклеотид» означает любой нуклеотид или аналог нуклеотида, который может быть представлен с помощью любого символа, определённого в Приложении 1 (см. раздел 1, Таблица 1), где нуклеотид или аналог нуклеотида содержат:

- (i) основную часть, выбранную из:

- (1) 2'-дезоксирибоза 5'-монофосфата (основной цепи дезоксирибонуклеотида) или рибоза 5'-монофосфата (основной цепи рибонуклеотида); или
- (2) аналог 2'-дезоксирибозо-5'-монофосфата или рибозо 5'-монофосфата, который при образовании основной цепи аналога нуклеиновой кислоты приводит к расположению нуклеотидных оснований, которые имитируют расположение нуклеотидов в нуклеиновых кислотах, содержащих 2'-дезоксирибозо-5'-монофосфата или рибозо-5'-монофосфата, где аналог нуклеиновой кислоты способен к сдвигу оснований с комплементарной нуклеиновой кислотой; примеры нуклеотидных аналогов включают аминокислоты, такие как в пептидных нуклеиновых кислотах, молекулы гликоля, как в гликоль нуклеиновых кислотах, молекулы треофуранозилового сахара, как в треозо-нуклеиновых кислотах, морфолиновые кольца и фосфородиамидатные группы, как в морфолинах, и молекулы циклогексенила, как в циклогексенильных нуклеиновых кислотах.

и

(ii) основной частью является либо:

(1) присоединенное к нуклеотидному основанию, включая модифицированное или синтетическое пуриновое или пиримидиновое нуклеотидное основание; или

(2) отсутствие пуриновой или пиримидиновой нуклеотидной группы, когда нуклеотид является частью нуклеотидной последовательности, называемой «AP-site» или «abasic site» (участок ДНК, лишенный азотистых оснований).

h) «остаток» означает любой отдельный нуклеотид или аминокислоту, или их соответствующие аналоги в последовательности.

i) «идентификационный номер последовательности» представляет собой уникальный номер (целое число), присваиваемый каждой последовательности в перечне последовательностей.

j) «перечень последовательностей» означает как часть описания поданной заявки на патент, так и документ, поданный позднее заявки, который включает раскрытие нуклеотидной и/или аминокислотной последовательности(ей), а также другую имеющуюся информацию, как предусмотрено настоящим Стандартом.

к) «специально определённые» означает любые нуклеотиды, кроме тех, которые представлены с помощью "n", и любые аминокислоты, кроме тех, которые представлены с помощью "X", из перечисленных в Приложении 1 (см. раздел 1, Таблица 1 и раздел 3, Таблица 3 соответственно).

l) «неизвестный» нуклеотид или аминокислота означает, что отдельный нуклеотид или аминокислота представлены, но их идентификационные данные неизвестны или нераскрыты.

4. Для целей настоящего Стандарта слово (слова):

(a) «может» относится к необязательному или допустимому методу, но не к требованию.

(b) «должен» относится к требованию; игнорирование этого требования приведет к несоблюдению Стандарта.

(c) «не должно» относится к запрету в Стандарте.

(d) «следует» относится к настоятельно рекомендованному подходу, но не к требованию.

(e) «не следует» относится к настоятельно не рекомендованному методу, но не к запрету.

Область применения

5. Настоящий Стандарт устанавливает требования к представлению перечня последовательностей нуклеотидов и аминокислот в патентных заявках.

6. Перечни последовательностей, соответствующие данному Стандарту (далее - перечни последовательностей), содержат часть общей информации и часть с данными

последовательностей. Перечень последовательностей должен быть представлен как единый файл в формате Расширяемого Языка Разметки (Extensible Markup Language, далее - XML), использующего определение типа документа (Document Type Definition, далее - DTD), представленные в Приложении 2. Часть общей информации с содержащейся в ней библиографической информацией служит исключительно для объединения её с перечнем последовательностей в патентной заявке, для которой этот перечень был представлен. Часть с данными последовательностей состоит из одного или более элементов данных последовательности, каждый из которых содержит информацию об одной последовательности. Элементы данных последовательности включают различные свойства (характеристики) ключей и квалификаторов последовательностей, на основе спецификаций Совместной Международной Базы Данных Нуклеотидных Последовательностей (International Nucleotide Sequence Database Collaboration, далее - INSDC) и Консорциума UniProt.

7. Для целей данного Стандарта последовательность, для которой необходим перечень последовательностей, раскрывается путём перечисления её остатков и может быть представлена:

а) неразветвлённой последовательностью или линейным участком разветвлённой последовательности, содержащей десять и более специально определённых нуклеотидов, где смежные нуклеотиды соединены либо:

(i) от 3' к 5' (или от 5' к 3') фосфодиэфирной связью; либо

(ii) любой химической связью, которая приводит к такому расположению соседних нуклеотидных оснований, которые имитируют расположение нуклеотидных оснований в нуклеиновых кислотах природного происхождения; или

б) неразветвлённой последовательностью или линейным участком разветвлённой последовательности, содержащей четыре и более специально определённых аминокислот, где смежные аминокислоты соединены пептидными связями.

8. Перечень последовательностей не должен включать никаких последовательностей с менее чем десятью специально определёнными нуклеотидами или с менее чем четырьмя специально определёнными аминокислотами, как последовательность со своим собственным присвоенным идентификационным номером последовательности.

Отсылки

9. Следующие Стандарты и документы являются релевантными по отношению к данному Стандарту:

Совместная Международная База
Данных Нуклеотидных
Последовательностей (INSDC)

<http://www.insdc.org/>;

Международный Стандарт ISO 639-1:2002

Коды для обозначения названий языков
Часть 1: Коды Alpha-2;

Консорциум UniProt:

<http://www.uniprot.org/>

W3C XML 1.0

<http://www.w3.org/>

Стандарт ВОИС ST.2	Стандартное представление календарных дат на опубликованных документах с использованием григорианского календаря Рекомендуемый Стандарт на двубуквенные коды для представления стран, административных единиц и межправительственных организаций; Рекомендуемые стандартные коды для идентификации различных видов патентных документов Стандарт по представлению перечней последовательностей нуклеотидов и аминокислот в патентных заявках
Стандарт ВОИС ST.3	
Стандарт ВОИС ST.16	
Стандарт ВОИС ST.25	

Представление последовательностей

10. Каждой последовательности, охваченной параграфом 7, должен быть присвоен отдельный идентификационный номер, включая последовательность, которая идентична участку более длинной последовательности. Идентификационный номер последовательности должен начинаться с 1 и последовательно увеличиваться на целое число. При отсутствии последовательности представляется идентификационный номер последовательности, т.е. на месте намеренно пропущенной последовательности должен использоваться «000» (см. раздел 58). Общее количество последовательностей должно быть указано в перечне последовательностей и должно равняться общему количеству идентификационных номеров последовательности, независимо от того, будет ли это некоторая последовательность или код «000».

Нуклеотидные последовательности

11. Нуклеотидная последовательность должна быть представлена единой цепочкой слева направо в направлении от 5'-конца к 3'-концу или в направлении слева направо, которое имитирует направление от 5'-конца к 3'-концу. Обозначения 5' и 3' или любые другие подобные обозначения не должны быть включены в последовательность. Нуклеотидная последовательность, состоящая из двух цепочек, раскрывается путём перечисления остатков обеих цепочек, которые должны быть представлены как:

a) единая последовательность или как две отдельных последовательности, каждой из которых присвоен свой собственный идентификационный номер последовательности, в тех случаях, когда две отдельные цепочки полностью комплементарны друг другу; или

b) две отдельные последовательности, каждой из которых присвоен свой собственный идентификационный номер последовательности, в тех случаях, когда две отдельные цепочки не полностью комплементарны друг другу.

12. Для целей настоящего Стандарта первый нуклеотид, представленный в последовательности, является остатком на позиции номер 1. Когда нуклеотидная последовательность имеет циклическую (кольцевую) конфигурацию, заявитель может выбрать нуклеотид с положением остатка номер 1. Нумерация непрерывна по всей последовательности в направлении от 5' к 3' или в направлении, которое имитирует направление от 5' до 3'. Последний номер позиции остатка должен быть равен числу нуклеотидов в последовательности.

13. Все нуклеотиды в последовательности должны быть представлены с использованием символов, определённых в Приложении 1 (см. Глава 1, Таблица 1). При этом используются только строчные буквы. Любой символ, используемый для представления нуклеотида, является эквивалентом только одного остатка.

14. Символ “t” будет истолкован как тимин в ДНК и урацил в РНК. Урацил в ДНК или тимин в РНК считается модифицированным нуклеотидом и должен быть в дальнейшем описан в таблице характеристик, как представлено в параграфе 19.

15. В случае, когда присвоен неоднозначный символ (представляющий два и более альтернативных нуклеотида), должен быть использован наиболее ограничивающий символ, как перечислено в Приложении 1 (Глава 1, Таблица 1). Например, в случае если нуклеотидами данной позиции могут быть “a” или “g”, следует использовать скорее “r”, чем “n”. Символ “n” будет рассматриваться как “a” или “c” или “g” или “t/u”, если он используется без дополнительного описания, как того требуют параграфы 16 и 17 или 21. Символ “n” не должен использоваться для представления чего-либо, кроме нуклеотида. Символ “n” может использоваться для представления только одного модифицированного или “неизвестного” нуклеотида, вместе с последующим описанием в таблице характеристик, как предусмотрено в параграфах 16 и 17 или 21. Для представления вариантов последовательности, то есть альтернативы, удаления, вставки или замены, смотри параграфы 92-98.

16. Модифицированные нуклеотиды должны быть представлены в последовательности как соответствующие немодифицированные нуклеотиды, т.е. как “a”, “c”, “g” или “t” всякий раз, когда это возможно. Любой модифицированный нуклеотид в последовательности, который не может быть представлен иначе, как любой другой символ в Приложении 1 (см. Глава 1, Таблица 1), т.е., «другой» нуклеотид, такой как неприродные нуклеотиды, должен быть представлен символом “n.” В тех случаях, когда символ “n” используется для представления модифицированного нуклеотида, он является эквивалентом только одного основания.

17. Модифицированный нуклеотид должен быть в дальнейшем описан в таблице свойств (см. параграф 60 и последующие) с использованием специального ключа “modified_base” и обязательного квалификатора “mod_base” в сочетании с единственной аббревиатурой из Приложения 1 (см. Глава 2, Таблица 2) в качестве значения квалификатора; если аббревиатурой является “OTHER” (“ДРУГОЙ”), полное несокращённое название модифицированного нуклеотида должно быть приведено как значение квалификатора “note”. Для перечисления альтернативных модифицированных нуклеотидов значение квалификатора «OTHER» может использоваться в сочетании с дополнительным квалификатором “note” («примечания») (см. параграфы 95 и 96). Аббревиатуры (или полные названия), приведенные в Приложении 1 (см. Глава 2, Таблица 2) не должны быть использованы в самой последовательности.

18. Нуклеотидная последовательность, включающая один или несколько участков следующих друг за другом модифицированных нуклеотидов, которые имеют одну и ту же основную часть (см. параграф 3 (g) (i) (2)), должна быть дополнительно описана в таблице характеристик, как это предусмотрено в параграфе 17. Модифицированные нуклеотиды каждого такого участка могут быть совместно описаны в одном элементе INSDFeature, как это предусмотрено в параграфе 22. В качестве значения в квалификаторе “note” («примечания») должно быть указано наиболее ограничивающее несокращённое химическое название, которое охватывает все модифицированные нуклеотиды в пределах участка или перечень химических названий всех нуклеотидов в диапазоне. Например, последовательность нуклеиновой кислоты гликоля, содержащая нуклеотиды нуклеинов «a», «c», «g» или «t», может быть описана в квалификаторе “note” как «2,3-дигидроксипропилнуклеозиды».

Альтернативно, такая же последовательность может быть описана в квалификаторе “note” как «2,3-дигидроксипропиладенин, 2,3-дигидроксипропилтимин, 2,3-дигидроксипропилгуанин или 2,3-дигидроксипропилцитозин». Если отдельный модифицированный нуклеотид на участке включает дополнительную модификацию, то модифицированный нуклеотид также должен быть дополнительно описан в таблице характеристик, как это предусмотрено в параграфе 17.

19. Урацил в ДНК или тимин в РНК считаются модифицированными нуклеотидами и должны быть представлены в последовательности как “t” и далее описаны в таблице функций с помощью специального ключа “modified_base”, квалификатора “mod_base” с “OTHER” как значением квалификатора и квалификатора “note” с “урацилом” или “тиминном”, соответственно, как значение квалификатора.

20. Следующие примеры иллюстрируют модифицированные нуклеотиды, представленные в соответствии с параграфами с 16 по 18 выше:

Пример 1: Модифицированный нуклеотид, использующий аббревиатуру из Приложения 1 (см. Глава 2, Таблица 2)

```
<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>
```

Пример 2: Модифицированный нуклеотид “xanthine”, использующий “OTHER” из Приложения 1 (см. Глава 2, Таблица 2)

```
<INSDFeature>
  <INSDFeature_key>modified_base</INSDFeature_key>
  <INSDFeature_location>4</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>
```

Пример 3. Нуклеотидная последовательность, состоящая из модифицированных нуклеотидов, охватываемых параграфом 3 (g) (i) (2), с двумя отдельными нуклеотидами, которые включают в себя дополнительную модификацию

```

<INSDFeature>
  <INSDFeature_key>modified_base</INSDFeature_key>
  <INSDFeature_location>1..954</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</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>

<INSDFeature>
  <INSDFeature_key>modified_base</INSDFeature_key>
  <INSDFeature_location>439</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>mod_base</INSDQualifier_name>
      <INSDQualifier_value>i</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>modified_base</INSDFeature_key>
  <INSDFeature_location>684</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>

```

21. Любой “unknown” («неизвестный») нуклеотид должен быть представлен в последовательности как “n”. “Unknown” («неизвестный») нуклеотид должен быть в дальнейшем описан в таблице свойств (см. раздел 60 и последующих) с помощью функционального ключа “unsure” (“неопределённый”). Символ “n” является эквивалентом только одного остатка.

22. Участок, содержащий известное число смежных остатков “a”, “c”, “g”, “t” или “n”, для которых применяют одинаковые описания, может быть описан с использованием одного элемента INSDFeature с синтаксисом “x.y” в качестве дескриптора местонахождения в элементе INSDFeature_location (см. параграфы с 64 - 71). Для представления вариантов последовательности, т.е. удаления, добавления или замещения, см. параграфы 92 - 98.

23. Следующий пример иллюстрирует представление области модифицированных нуклеотидов, для которых используют одинаковые описания, в соответствии с разделом 22 выше:

```
<INSDFeature>
  <INSDFeature_key>modified_base</INSDFeature_key>
  <INSDFeature_location>358..485</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>isoguanine</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
```

Аминокислотные последовательности

24. Аминокислоты в аминокислотной последовательности должны быть представлены в аминокислотном направлении слева направо. Амино и карбокси группы не должны быть представлены в последовательности.

25. Для целей настоящего Стандарта, первая аминокислота в последовательности является остатком на позиции номер 1, включая аминокислоты, предшествующие зрелому белку, например, пред-последовательности, про-последовательности, пред-про-последовательности и сигнальные последовательности. Когда аминокислотная последовательность имеет циклическую (кольцевую) конфигурацию, заявитель может выбрать аминокислоту с положением остатка номер 1. Нумерация непрерывна по всей последовательности в аминокислотном направлении.

26. Все аминокислоты в последовательности должны быть представлены с использованием символов, определённых в Приложении 1 (см. Главу 3, Таблицу 3). Должны использоваться только заглавные буквы. Любой символ, используемый для представления аминокислоты, является эквивалентом только одного остатка.

27. В случае, если уместным является неоднозначный символ (представляющий, как альтернативу, две и более аминокислоты), следует использовать наиболее ограничивающий символ. Например, если аминокислота на данной позиции может быть аспарагиновой кислотой или аспарагином, предпочтительнее использовать символ "B" нежели символ "X". Символ "X" будет истолкован как один из символов "A", "R", "N", "D", "C", "Q", "E", "G", "H", "I", "L", "K", "M", "F", "P", "O", "S", "U", "T", "W", "Y", или "V", за исключением случая использования последующего описания в таблице характеристик, как представлено в параграфах 29-31 или 32-33. Символ "X" не должен использоваться для представления чего-либо иного, кроме аминокислоты. Символ "X" может быть использован для представления только одной аминокислоты, вместе с последующим описанием в таблице характеристик, как предусмотрено в параграфах 29-31 или 32-33. Для представления вариантов последовательности, т.е. альтернатив, удаления, добавления или замещения, смотри параграфы 92 - 98.

28. Раскрытые аминокислотные последовательности, разделённые символами внутреннего разграничения, например, "Ter" («конец») или звездочкой "*" или точкой "." или пробелом, должны быть включены как отдельные последовательности для каждой аминокислотной последовательности, которая содержит, по крайней мере, четыре специально определённые аминокислоты, и упомянуты в параграфе 7. Каждой такой отдельной последовательности должен быть присвоен собственный идентификационный номер последовательности. Разделительные символы и пробелы не должны быть включены в последовательности

перечня последовательностей (см. параграф 57).

29. Когда это возможно, модифицированные аминокислоты, включая D-аминокислоты, должны быть представлены в последовательности, как соответствующие немодифицированные аминокислоты. Любая модифицированная аминокислота в последовательности, которая не может быть представлена любым символом из Приложения 1 (см. Глава 3, Таблица 3), т.е. «другая» аминокислота, должна быть представлена как “X”. Символ “X” является эквивалентом только одного остатка.

30. Модифицированная аминокислота должна быть далее описана в таблице характеристик (см. параграф 60 и последующие). Там, где это применимо, следует использовать функциональные ключи «CARBOHYD» или «LIPID» вместе с квалификатором “NOTE” («ПРИМЕЧАНИЕ»). Функциональный ключ “MOD_RES” следует использовать для других посттрансляционно модифицированных аминокислот совместно с квалификатором “NOTE”; в других случаях функциональный ключ “SITE” следует использовать совместно с квалификатором “NOTE”. Значение квалификатора “NOTE” должно быть или аббревиатурой (сокращением), определённой в Приложении 1 (см. Глава 4, Таблица 4), или полным, несокращённым названием модифицированной аминокислоты. Аббревиатуры, определённые в Таблице 4, о которых речь идёт выше, или полные, несокращённые названия не должны использоваться в самой последовательности.

31. Следующие примеры иллюстрируют представление модифицированных аминокислот в соответствии с параграфом 30 выше:

Пример 1: Посттрансляционно модифицированная аминокислота

```
<INSDFeature>
  <INSDFeature_key>MOD_RES</INSDFeature_key>
  <INSDFeature_location>3</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>3Hyp</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
```

Пример 2: Непосттрансляционно модифицированная аминокислота

```
<INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>3</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>Orn</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
```

Пример 3: D-аминокислота

```
<INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>9</INSDFeature_location>
  <INSDFeature_qual>
```

```
<INSDQualifier>
  <INSDQualifier_name>NOTE</INSDQualifier_name>
  <INSDQualifier_value>D-Arginine</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_qual>
</INSDFeature>
```

32. Любая "unknown" («неизвестная») аминокислота, должна быть представлена как символ "X" в последовательности. "Unknown" («неизвестная») аминокислота, указанная как "X" должна быть в дальнейшем описана в таблице характеристик (см. параграф 60 и последующие) с помощью функционального ключа "UNSURE" и возможно (по выбору) квалификатора "NOTE". Символ "X" является эквивалентом только одного остатка.

33. Следующие примеры иллюстрируют представление "unknown" («неизвестной») или "other" («другой») аминокислоты в соответствии с параграфом 32 выше:

```
<INSDFeature>
  <INSDFeature_key>UNSURE</INSDFeature_key>
  <INSDFeature_location>3</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>A or V</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
```

34. Участки, содержащие известное число смежных остатков "X", для которых используют одинаковые описания, могут быть совместно описаны с помощью синтаксиса "x..y", как описание местонахождения в элементе INSDFeature_location (см. параграфы 64-70). Для представления вариантов последовательности, т.е. удаления, добавления или замещения, см. параграфы 93 - 98.

Представление особых ситуаций

35. Последовательность, раскрытая путём перечисления её остатков, которые построены как единая, непрерывная последовательность из одного или более несмежных сегментов более длинной последовательности или сегментов из различных последовательностей, должна быть включена в перечень последовательностей с присвоенным собственным идентификационным номером последовательности.

36. Последовательность, которая содержит участки специально определённых остатков, отделённых одной или более областями с конфигурациями остатков "n" или "X" (см. параграфы 15 и 27, соответственно), где раскрыто точное число "n" или "X" остатков на каждом участке, должна быть включена в перечень последовательностей как одна последовательность с присвоенным собственным идентификационным номером последовательности.

37. Последовательность, которая содержит участки специально определённых остатков, отделённых одним или более интервалами неизвестных или нераскрытых остатков, не должна быть представлена в перечне последовательностей как отдельная последовательность. Каждый участок специально определённых остатков, который охвачен параграфом 7, должен быть включен в перечень последовательностей в виде отдельной последовательности с присвоенным собственным идентификационным номером последовательности.

Структура перечня последовательностей в XML

38. На основании параграфа 6 выше, файл перечня последовательностей в XML в соответствии с этим Стандартом состоит из:

- (a) части общей информации, которая содержит информацию о патентной заявке, к которой относится перечень последовательностей; и
- (b) части с данными последовательности, которая содержит один или более элементов данных последовательности, каждый из которых, в свою очередь, содержит информацию об одной последовательности.

Пример перечня последовательностей представлен в Приложении 3.

39. Перечень последовательностей должен быть представлен в XML 1.0 с использованием DTD, представленных в Приложении 2 «Определение типа документа для перечня последовательностей».

- (a) Первая строка XML файла должна содержать XML объявление:

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

- (b) Вторая строка XML файла должна содержать объявление типа документа (DOCTYPE):

```
<!DOCTYPE ST26SequenceListing PUBLIC "-//WIPO//DTD Sequence Listing 1.0//EN"  
"ST26SequenceListing_V1_0.dtd">.
```

40. Весь электронный перечень последовательностей должен содержаться в одном файле. Файл должен быть закодирован с использованием уникода UTF-8 со следующими ограничениями:

- a) информация, содержащаяся в элементах ApplicantName, InventorName и InventionTitle части общей информации, может быть составлена из любых символов Unicode за исключением зарезервированных символов, которые должны быть представлены как определено в разделе 41; и

- b) информация, содержащаяся в остальных элементах части общей информации и всех элементах части с данными последовательностей, частично должна быть составлена из печатных символов (включая пробелы) кодовой таблицы Unicode Basic Latin, за исключением зарезервированных символов, которые должны быть представлены как определено в параграфе 41, (т.е. ограничена кодовыми точками Unicode 0020, 0021, с 0023 по 0026, с 0028 по 003B, 003D, и с 003F по 007E, смотри Приложение 4), и только разрешёнными символами, заранее определёнными элементами, оговоренными в параграфе 41.

41. В случае перечня последовательностей в XML, перечисленные зарезервированные символы должны быть заменены соответствующими, заранее определёнными элементами, когда они используются в значении признака или содержании элемента:

Зарезервированные символы	Заранее определённые элементы
<	<
>	>
&	&
"	"
'	'

Смотри раздел 71 в качестве примера.

42. Все обязательные элементы должны быть заполнены (за исключением предусмотренного в разделе 58 для намеренно пропущенной последовательности). Необязательные элементы, для которых недоступно содержание, не должны появляться в случае XML (за исключением случаев, предусмотренных в параграфе 95, для представления удаления в последовательности в качестве значения квалификатора “replace” («замена»)).

Корневой (основной) элемент

43. В соответствии с данным Стандартом корневым элементом в XML является элемент ST26SequenceListing, имеющий следующие признаки:

Признак	Описание	Обязательно/Необязательно
dtdVersion	Версия DTD, используемая для создания этого файла в формате “V#_#”, например. “V1_0”.	Обязательно
fileName	Файл с названием перечня последовательностей.	Необязательно
softwareName	Название программного обеспечения, создавшего этот файл.	Необязательно
softwareVersion	Версия программного обеспечения, создавшего этот файл.	Необязательно
productionDate	Дата создания файла перечня последовательностей (format “CCYY-MM-DD”).	Необязательно

44. Следующий пример иллюстрирует корневой элемент ST26SequenceListing и его признаки в XML, согласно параграфа 43 выше:

```
<ST26SequenceListing dtdVersion="V1_0" fileName="US11_405455_SEQ1.xml"
softwareName="SQL-software-name" softwareVersion="1.0"
productionDate="2006-05-10">
    {...}*
</ST26SequenceListing>
```

*{...} представляет часть общей информации и часть данных последовательности, которые не включены в этот пример.

Часть общей информации

45. Элементы части общей информации относятся к информации о патентной заявке и содержат следующие понятия:

Элемент	Описание	Обязательно/ Необязательно
ApplicationIdentification	Идентификация заявки, для которой представлен перечень последовательностей	Обязательно, если перечень последовательностей предоставлен в любое время после присвоения номера заявки
ApplicationIdentification состоит из:		
IPOfficeCode	Код ведомства подачи в соответствии со Стандартом ВОИС ST.3	Обязательно
ApplicationNumberText	Идентификация заявки, как установлено ведомством подачи (например, номер заявки PCT/IB2013/99999)	Обязательно
FilingDate	Дата подачи патентной заявки, для которой представлен перечень последовательностей (WIPO ST.2 формат "CCYY-MM-DD", используется 4-значное обозначение календарного года, 2-значное календарного месяца и 2-значное дня в течение календарного месяца, например, 2015-01-31)	Обязательно, если перечень последовательностей предоставлен в любое время после установления даты подачи
ApplicantFileReference	Отдельный уникальный идентификатор, присвоенный заявителю для идентификации отдельной заявки, введенный в символах, определенных в параграфе 40 (b)	Обязательно, если перечень последовательностей предоставлен в любое время до присвоения номера заявки; в противном случае необязательно

Элемент	Описание	Обязательно/ Необязательно
EarliestPriorityApplicationIdentification	Идентификация наиболее ранней приоритетной заявки (также содержащей IPOfficeCode, ApplicationNumberText и FilingDate, смотри ApplicationIdentification выше)	Обязательно, если испрашивается приоритет
ApplicantName	Имя первого указанного заявителя, введённое в символах, определённых в параграфе 40 (a). Данный элемент включает обязательный атрибут languageCode как определено в параграфе 47.	Обязательно
ApplicantNameLatin	Если ApplicantName введено в символах иных, чем это определено в параграфе 40 b), перевод или транслитерация имени первого указанного заявителя должны быть также введены в символах, определённых в параграфе 40 b)	Обязательно, если ApplicantName содержит нелатинские символы
InventorName	Имя первого указанного изобретателя, введённое в символах, определённых в параграфе 40 (a)	Необязательно
InventorNameLatin	Если InventorName введено в символах иных, чем это определено в параграфе 40 b), перевод или транслитерация имени первого указанного изобретателя могут быть также введены в символах, определённых в параграфе 40 b)	Необязательно

Элемент	Описание	Обязательно/ Необязательно
InventionTitle	Название изобретения, введённое в символах, определённых в параграфе 40 (a) на языке подачи. Перевод названия изобретения на дополнительные языки может быть введен в символах, определённых в параграфе 40 (a) используя дополнительный элемент InventionTitle. Данный элемент включает обязательный признак languageCode, как указано в параграфе 48. Предпочтительно, название изобретения содержит от двух до семи слов.	Обязательно на языке подачи. Необязательно для дополнительных языков.
SequenceTotalQuantity	Общее количество всех последовательностей в перечне, в том числе намеренно пропущенные последовательности (также известные как пустые последовательности) (см. параграф 10).	Обязательно

46. Следующие примеры иллюстрируют представление части общей информации о перечне последовательностей, согласно разделу 45 выше:

Пример 1: перечень последовательностей, представленный до идентификации заявки и даты подачи

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ST26SequenceListing PUBLIC "-//WIPO//DTD Sequence Listing
1.0//EN" "ST26SequenceListing_V1_0.dtd">
<ST26SequenceListing dtdVersion="V1_0" fileName="Invention_SEQ1.xml"
softwareName="SEQ1-software-name" softwareVersion="1.0"
productionDate="2015-05-10">
  <ApplicantFileReference>AB123</ApplicantFileReference>
  <EarliestPriorityApplicationIdentification>
    <IPOfficeCode>IB</IPOfficeCode>
    <ApplicationNumberText>PCT/IB2013/099999</ApplicationNumberText>
    <FilingDate>2014-07-10</FilingDate>
  </EarliestPriorityApplicationIdentification>
  <ApplicantName languageCode="en">GENOS Co., Inc.</ApplicantName>
  <InventorName languageCode="en">Keiko Nakamura</InventorName>
  <InventionTitle languageCode="en">SIGNAL RECOGNITION PARTICLE RNA AND
PROTEINS</InventionTitle>
```

```
<SequenceTotalQuantity>9</SequenceTotalQuantity>
<SequenceData sequenceIDNumber="1"> {...}* </SequenceData>
<SequenceData sequenceIDNumber="2"> {...} </SequenceData>
<SequenceData sequenceIDNumber="3"> {...} </SequenceData>
<SequenceData sequenceIDNumber="4"> {...} </SequenceData>
<SequenceData sequenceIDNumber="5"> {...} </SequenceData>
<SequenceData sequenceIDNumber="6"> {...} </SequenceData>
<SequenceData sequenceIDNumber="7"> {...} </SequenceData>
<SequenceData sequenceIDNumber="8"> {...} </SequenceData>
<SequenceData sequenceIDNumber="9"> {...} </SequenceData>
</ST26SequenceListing>
```

*{...} представляет релевантную информацию для каждой последовательности, которая не включена в данный пример.

Пример 2: перечень последовательностей, представленный после идентификации заявки и даты подачи

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ST26SequenceListing PUBLIC "-//WIPO//DTD Sequence Listing
1.0//EN" "ST26SequenceListing_V1_0.dtd">
<ST26SequenceListing dtdVersion="1_0" fileName="Invention_SEQI.xml"
softwareName="SEQI-software-name" softwareVersion="1.0"
productionDate="2015-05-10">
  <ApplicationIdentification>
    <IPOfficeCode>US</IPOfficeCode>
    <ApplicationNumberText>14/999,999</ApplicationNumberText>
    <FilingDate>2015-01-05</FilingDate>
  </ApplicationIdentification>
  <ApplicantFileReference>AB123</ApplicantFileReference>
  <EarliestPriorityApplicationIdentification>
    <IPOfficeCode>IB</IPOfficeCode>
    <ApplicationNumberText>PCT/IB2014/099999</ApplicationNumberText>
    <FilingDate>2014-07-10</FilingDate>
  </EarliestPriorityApplicationIdentification>
  <ApplicantName languageCode="en">GENOS Co., Inc.</ApplicantName>
  <InventorName languageCode="en">Keiko Nakamura</InventorName>
  <InventionTitle languageCode="en">SIGNAL RECOGNITION PARTICLE RNA AND
PROTEINS</InventionTitle>
  <SequenceTotalQuantity>9</SequenceTotalQuantity>
  <SequenceData sequenceIDNumber="1"> {...}* </SequenceData>
  <SequenceData sequenceIDNumber="2"> {...} </SequenceData>
  <SequenceData sequenceIDNumber="3"> {...} </SequenceData>
  <SequenceData sequenceIDNumber="4"> {...} </SequenceData>
  <SequenceData sequenceIDNumber="5"> {...} </SequenceData>
  <SequenceData sequenceIDNumber="6"> {...} </SequenceData>
  <SequenceData sequenceIDNumber="7"> {...} </SequenceData>
  <SequenceData sequenceIDNumber="8"> {...} </SequenceData>
  <SequenceData sequenceIDNumber="9"> {...} </SequenceData>
</ST26SequenceListing>
```

*{...} представляет релевантную информацию для каждой последовательности, которая не включена в данный пример.

47. Название заявителя и, необязательно, имя изобретателя должны быть указаны в элементах `ApplicantName` и `InventorName`, соответственно, так как они в основном приводятся на языке, на котором подана заявка. Для каждого элемента должен быть указан соответствующий языковой код (см. отсылку в параграфе 9 к ISO 639-1:2002) в параметре `languageCode`. Когда указание названия заявителя содержит символы иные, чем это определено в параграфе 40 (b), транслитерация или перевод названия заявителя должны быть также указаны в символах латинского алфавита в элементе `ApplicantNameLatin`. Когда указание имени изобретателя содержит символы, отличные от символов латинского алфавита, транслитерация или перевод имени изобретателя могут быть также указаны в символах латинского алфавита в элементе `InventorNameLatin`.

48. Название изобретения должно быть указано в элементе `InventionTitle` на языке подачи и может быть также указано на других языках, используя для этого элементы `InventionTitle` (см. таблицу в параграфе 45). Соответствующий языковой код (см. отсылку в параграфе 9 к ISO 639-1:2002) должен быть указан в параметре `languageCode` элемента.

49. Следующий пример иллюстрирует представление имени и названия изобретения согласно разделам 47 и 48 выше:

Пример: Название заявителя и имя изобретателя каждое представлено в буквах японского и латинского алфавита, и название изобретения, представленное в буквах японского, английского и французского алфавита

```
<ApplicantName languageCode="ja">出願製薬株式会社</ApplicantName>
<ApplicantNameLatin>Shutsugan Pharmaceuticals Kabushiki
Kaisha</ApplicantNameLatin>
<InventorName languageCode="ja">特許太郎</InventorName>
<InventorNameLatin>Taro Tokkyo</InventorNameLatin>
<InventionTitle languageCode="ja">efg タンパク質をコードするマウス abcd-1 遺伝子
</InventionTitle>
<InventionTitle languageCode="en">Mus musculus abcd-1 gene for efg protein
</InventionTitle>
<InventionTitle languageCode="fr">Gène abcd-1 de Mus musculus pour
protéine efg</InventionTitle>
```

Часть с данными последовательности

50. Часть с данными последовательности должна состоять из одного или более элементов `SequenceData`, каждый элемент содержит информацию об одной последовательности.

51. Каждый элемент `SequenceData` должен иметь обязательный признак `sequenceIDNumber`, в котором содержится идентификационный номер (см. параграф 10) для каждой последовательности. Например:

```
<SequenceData sequenceIDNumber="1">
```

52. Элемент `SequenceData` содержит зависимый элемент `INSDSeq`, который в свою очередь содержит следующие зависимые элементы:

Элемент	Описание	Обязательно/Не включено	
		Последовательности	Намеренно пропущенные последовательности
INSDSeq_length	Длина последовательности	Обязательно	Обязательно без значения
INSDSeq_moltype	Тип молекулы	Обязательно	Обязательно без значения
INSDSeq_division	Указание, что последовательность относится к патентной заявке	Обязательно со значением "PAT"	Обязательно без значения
INSDSeq_feature-table	Список аннотаций последовательностей	Обязательно	Не должны быть включены
INSDSeq_sequence	Последовательность	Обязательно	Обязательно со значением "000"

53. Элемент `INSDSeq_length` должен показывать число нуклеотидов или аминокислот последовательности, содержащейся в элементе `INSDSeq_sequence`. Например:

```
<INSDSeq_length>8</INSDSeq_length>
```

54. Элемент `INSDSeq_moltype` должен показывать тип молекулы, которая была представлена. Для нуклеотидных последовательностей, включая последовательности аналогов нуклеотидов, тип молекулы должен быть указан как ДНК или РНК. Для аминокислотных последовательностей тип молекулы должен быть указан как АА. (Данный элемент отличен от квалификаторов "mol_type" и "MOL_TYPE", рассмотренных в параграфах 55 и 84). Например:

```
<INSDSeq_moltype>AA</INSDSeq_moltype>
```

55. В тех случаях, когда нуклеотидная последовательность содержит как фрагмент ДНК, так и фрагмент РНК, значением для `INSDSeq_moltype` должен быть "ДНК.". Комбинированная молекула ДНК / РНК должна быть далее описана в таблице характеристик с помощью функционального ключа "source" и обязательного квалификатора "organism" со значением "synthetic construct" и обязательного квалификатора "mol_type" со значением "other ДНК.". Каждый ДНК и РНК фрагмент комбинированной ДНК/РНК молекулы должен быть в дальнейшем описан с функциональным ключом "misc_feature" и квалификатором "note", который указывает, является ли фрагмент ДНК или РНК.

56. Следующий пример иллюстрирует описание нуклеотидной последовательности, содержащей фрагменты и ДНК и РНК., согласно параграфу 55 выше.

```
<INSDSeq>
  <INSDSeq_length>120</INSDSeq_length>
  <INSDSeq_moltype>DNA</INSDSeq_moltype>
  <INSDSeq_division>PAT</INSDSeq_division>
  <INSDSeq_feature-table>
    <INSDFeature>
      <INSDFeature_key>source</INSDFeature_key>
      <INSDFeature_location>1..120</INSDFeature_location>
      <INSDFeature_qual>
        <INSDQualifier>
```

```

    <INSDQualifier_name>organism</INSDQualifier_name>
    <INSDQualifier_value>synthetic construct</INSDQualifier_value>
  </INSDQualifier>
  <INSDQualifier>
    <INSDQualifier_name>mol_type</INSDQualifier_name>
    <INSDQualifier_value>other DNA</INSDQualifier_value>
  </INSDQualifier>
</INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>misc_feature</INSDFeature_key>
  <INSDFeature_location>1..60</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>note</INSDQualifier_name>
      <INSDQualifier_value>DNA fragment</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>misc_feature</INSDFeature_key>
  <INSDFeature_location>61..120</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>note</INSDQualifier_name>
      <INSDQualifier_value>RNA fragment</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>
  cgaccacgcgtccgaggaaccaaccatcacgtttgaggacttcgtgaaggaattggataataaccgctccctacca
  aaatggcgagcgccgactcattgctcctcgtagcgtagcgaggc
</INSDSeq_sequence>
</INSDSeq>

```

57. Элемент `INSDSeq_sequence` должен раскрывать последовательность. В последовательность должны быть включены только соответствующие символы, указанные в Приложении 1 (см. Глава 1, Таблица 1 и Глава 3, Таблица 3). Последовательность не должна включать цифры, знаки препинания или пробелы.

58. Намеренно пропущенные последовательности должны быть включены в перечень последовательностей и представлены следующим образом:

- a) элемент `SequenceData` и его признак `sequenceIDNumber` с идентификационным номером последовательности для пропущенной последовательности в качестве его значения;
- b) элементы `INSDSeq_length`, `INSDSeq_moltype`, `INSDSeq_division` представлены, но без значения;
- c) элемент `INSDSeq_feature-table` не должен быть включен, и
- d) элемент `INSDSeq_sequence` с цепочкой "000" в качестве значения.

59. Следующий пример иллюстрирует представление намеренно пропущенной последовательности, согласно параграфу 58 выше:

```
<SequenceData sequenceIDNumber="3">
  <INSDSeq>
    <INSDSeq_length/>
    <INSDSeq_moltype/>
    <INSDSeq_division/>
    <INSDSeq_sequence>000</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
```

Таблица характеристик

60. Таблица характеристик содержит информацию о месте и роли различных областей определённой последовательности. Таблица характеристик необходима для каждой последовательности, за исключением любой намеренно пропущенной последовательности, в этом случае она не приводится. Таблица характеристик содержится в элементе `INSDSeq_feature-table`, который включает один или более элементов `INSDFeature`.

61. Каждый элемент `INSDFeature` описывает один признак и содержит следующие зависимые элементы:

Элемент	Описание	Обязательно/Необязательно
<code>INSDFeature_key</code>	Слово или аббревиатура, указывающие признак	Обязательно
<code>INSDFeature_location</code>	Участок последовательности, соответствующий признаку	Обязательно
<code>INSDFeature_qual</code>	Квалификатор, содержащий вспомогательную информацию о признаке	Обязательно, где функциональный ключ требует одного или более квалификаторов, например источник; в противном случае необязательно

Функциональные ключи

62. Приложение 1 содержит единый перечень функциональных ключей, которые должны использоваться в соответствии с данным Стандартом, совместно с единым перечнем соответствующих квалификаторов и указанием, относится ли данный квалификатор к обязательным или необязательным. Единый перечень функциональных ключей для нуклеотидных последовательностей определён в Главе 5 Приложения 1, а для аминокислотных последовательностей – в Главе 7 Приложения 1.

Обязательные функциональные ключи

63. Функциональный ключ "source" является обязательным для всех нуклеотидных последовательностей, а функциональный ключ "SOURCE" является обязательным для всех аминокислотных последовательностей, за исключением любых намеренно пропущенных последовательностей. Каждая последовательность должна иметь единственный функциональный ключ "source" или "SOURCE", охватывающий всю последовательность. Когда

последовательность образуется из нескольких источников, данные источники могут быть в дальнейшем описаны в таблице характеристик с помощью функционального ключа “misc_feature” и квалификатора “note” для нуклеотидных последовательностей и функционального ключа “REGION” и квалификатора “NOTE” для аминокислотных последовательностей.

Местоположение признака

64. Обязательный элемент `INSDFeature_location` должен содержать, по крайней мере, одно описание местоположения, которое определяет место или участок, соответствующий признаку последовательности в элементе `INSDSeq_sequence`, и может содержать один или более оператор(ов) местоположения (см. параграфы 67-70).

65. Описанием местоположения может быть номер единичного остатка, место между номерами двух соседних остатков, участок ограниченный диапазоном смежных остатков, или место, или участок, который оказывается за пределами определённого остатка или диапазона остатков. Повторяющиеся описания местоположений могут быть использованы в соединении с оператором местоположения, когда признаку соответствуют дискретные места или области последовательности (смотри параграфы 67-70). Описание местоположения в элементе `INSDSeq_sequence` не должно включать нумерацию остатков за пределами последовательности.

66. Синтаксис для каждого типа описания местоположения указан в таблице ниже, где *x* и *y* являются номерами остатков, указанных как неотрицательные целые числа, не больше, чем длина последовательности в элементе `INSDSeq_sequence`, а *x* меньше, чем *y*.

Тип дескриптора местоположения	Синтаксис	Описание
Номер единичного остатка	<i>x</i>	Указывает на единичный остаток в последовательности
Номера остатков, ограниченных диапазоном последовательности	<i>x</i> . . <i>y</i>	Указывает на ограниченный непрерывный диапазон остатков, включая стартовый и конечный остатки.
Остатки перед первым или после последнего определённого номера остатка	< <i>x</i> > <i>x</i> < <i>x</i> . . <i>y</i> <i>x</i> . . > <i>y</i>	Указывает на область, включающую определённый остаток или диапазон остатков и определённый остаток, выходящий за рамки диапазона. Символы '<' и '>' могут быть использованы с единичным остатком или для указания возможного выхода из диапазона номера определённого остатка.
Место между номерами двух смежных остатков	<i>x</i> ^ <i>y</i>	Указывает на место между двумя смежными остатками, например место разложения эндонуклеазы, номера позиций для смежных остатков, отделённых каратом (^). Допустимыми форматами для этого дескриптора являются <i>x</i> ^ <i>x</i> + 1 (например 55 ^ 56), или, для циклических нуклеотидов <i>x</i> ^ 1, где “ <i>x</i> ” это полная длина молекулы, то есть 1000 ^ 1 для циклической молекулы с длиной 1000.

67. Оператор местоположения является префиксом для одного описания местоположения или комбинации дескрипторов местоположения, соответствующих одному, но характерному признаку, и определяет местоположение соответствующего признака в указанной последовательности или как построен признак. Перечень операторов местоположения с их определениями приводится ниже.

а) Оператор местоположения для нуклеотидов и аминокислот:

Синтаксическая структура местоположения	Описание местоположения
<code>join(location, location, ... location)</code>	Указатели местоположений объединены (сквозное размещение), чтобы сформировать непрерывную последовательность.
<code>order(location, location, ... location)</code>	Элементы находятся в определённом порядке, но ничто не предполагает целесообразности их соединения.

б) Оператор местоположения только для нуклеотидов:

Синтаксическая структура местоположения	Описание местоположения
<code>complement(location)</code>	Указывает, что признак расположен в дополнительной цепи диапазона последовательности, определённого указателем местоположения, когда считывается в 5' - 3' направлении или в направлении, которое имитирует 5' - 3' направление.

68. Соединение и порядок расположения операторов местоположения требуют, чтобы были представлены, по крайней мере, две запятые, разделяющие дескрипторы местоположения. Дескрипторы местоположения включаются между двумя соседними остатками, например `x^u`, и не должны использоваться в соединении или указании порядка расположения. Использование операторов соединения предполагает, что остатки, описываемые дескрипторами местоположения, физически контактируют в биологических процессах (например, экзоны, способствующие кодирующим свойствам участка)

69. Оператор местоположения “complement” может быть использован только для нуклеотидов. “Complement” может быть использован в сочетании либо с “join”, либо с “order” в одном и том же месте. Сочетания “join” и “order” в одном и том же месте не допускается.

70. Следующие примеры иллюстрируют местоположение признаков согласно параграфам 64 – 69 выше:

а) Местоположение для нуклеотидов и аминокислот:

Пример местоположения	Описание
467	Указывает на остаток 467 в последовательности.
123^124	Указывает на позицию между остатками 123 и 124.
340..565	Указывает на ограниченный непрерывный диапазон остатков, остатки 340 и 565.
<1	Указывает на признак местоположения перед первым остатком.
<345..500	Указывает на то, что точный нижний предел признака неизвестен. Размещение начинается с некоторого остатка, предшествующего 345 и продолжается до остатка 500 включительно.
<1..888	Указывает на то, что признак начинается перед первым остатком последовательности и продолжается до остатка 888 включительно.
1..>888	Указывает на то, что признак начинается с первого остатка последовательности и продолжается за пределами остатка 888.
join(12..78,134..202)	Указывает на то, что области с 12 по 78 и с 134 по 202 должны быть объединены в одну непрерывную последовательность.

b) Местоположение только для нуклеотидов:

Пример местоположения	Описание
complement(34..126)	Начинается с нуклеотида, комплементарного 126 основанию, и заканчивается на нуклеотиде, комплементарном нуклеотиду 34 (признак находится на цепи, комплементарной представленной).
complement(join(2691..4571,4918..5163))	Последовательность нуклеотидов с 2691 по 4571 и с 4918 по 5163, которая комплементарна объединенным сегментам (признак находится на цепи, комплементарной представленной)
join(complement(4918..5163),complement(2691..4571))	Объединение с 4918 по 5163 и с 2691 по 4571 оснований комплементарных цепей (признак находится на цепи, комплементарной представленной)

71. В перечне последовательностей XML символы "<" и ">" в дескрипторе местоположения должны быть заменены заранее определёнными элементами (см. параграф 41). Например:

```
Feature location "<1":
<INSDFeature_location>&lt;1</INSDFeature_location>

Feature location "1..>888":
<INSDFeature_location>1..&gt;888</INSDFeature_location>
```

Квалификаторы признака

72. Квалификаторы используются для получения информации о признаках в дополнение к той, которая даётся через функциональный ключ и местоположение признака. Есть три типа форматов для представления различных типов информации, предоставляемых квалификаторами, а именно:

- a) свободный текст (смотри параграфы 85 - 86);
- b) контролируемая лексика или числовые значения (например, номер или дата); и
- c) последовательности.

73. В Главе 6 Приложения 1 представлен единый перечень квалификаторов с определёнными значениями форматов, и если имеется, функциональным ключом для каждого нуклеотида и в главе 8 содержится единый перечень квалификаторов для каждого аминокислотного функционального ключа.

74. Любая последовательность, охваченная параграфом 7, которая представлена как значение квалификатора, должна быть включена в виде отдельной последовательности с присвоенным собственным идентификационным номером последовательности.

Обязательные квалификаторы признаков

75. Один обязательный функциональный ключ, например "source" для нуклеотидных последовательностей и "SOURCE" для аминокислотных последовательностей, требует двух обязательных квалификаторов, "organism" и "mol_type" для нуклеотидных последовательностей и "ORGANISM" и "MOL_TYPE" для аминокислотных последовательностей. Некоторые необязательные функциональные ключи также требуют обязательных квалификаторов.

Элементы квалификатора

76. Элемент `INSDFeature_qual` содержит один или более элементов `INSDQualifier`. Каждый элемент `INSDQualifier` представляет один квалификатор и состоит из двух зависимых элементов, указанных ниже:

Элемент	Описание	Обязательно/Необязательно
<code>INSDQualifier_name</code>	Название квалификатора (смотри Приложение 1, главы 6 и 8)	Обязательно
<code>INSDQualifier_value</code>	Значение квалификатора, если имеется, в определённом формате (смотри Приложение 1, главы 6 и 8)	Обязательно, когда точно определено (смотри Приложение 1, главы 6 и 8)

77. Квалификатор организма, т.е. “organism” для нуклеотидных последовательностей (смотри Приложение 1, глава 6) и “ORGANISM” для аминокислотных последовательностей (смотри Приложение 1, глава 8) должен раскрывать источник, например, отдельный организм или происхождение последовательности. Указанный организм должен быть выбран из систематической базы данных.

78. Если последовательность является природной и исходный организм имеет латинский род и обозначение вида, то обозначение должно быть использовано в качестве значения квалификатора. Предпочтительное английское общепринятое название может быть приведено с использованием квалификатора “note” для нуклеотидных последовательностей и квалификатора “NOTE” для аминокислотных последовательностей, но не должно быть использовано в значении квалификатора организма.

79. Следующие примеры иллюстрируют источник последовательности, согласно параграфам 77 – 78 выше:

Пример 1: Источник для нуклеотидной последовательности:

```
<INSDSeq_feature-table>
  <INSDFeature>
    <INSDFeature_key>source</INSDFeature_key>
    <INSDFeature_location>1..5164</INSDFeature_location>
    <INSDFeature_qual>
      <INSDQualifier>
        <INSDQualifier_name>organism</INSDQualifier_name>
        <INSDQualifier_value>Solanum lycopersicum</INSDQualifier_value>
      </INSDQualifier>
      <INSDQualifier>
        <INSDQualifier_name>note</INSDQualifier_name>
        <INSDQualifier_value>common name: tomato</INSDQualifier_value>
      </INSDQualifier>
      <INSDQualifier>
        <INSDQualifier_name>mol_type</INSDQualifier_name>
        <INSDQualifier_value>genomic DNA</INSDQualifier_value>
      </INSDQualifier>
    </INSDFeature_qual>
  </INSDFeature>
</INSDSeq_feature-table>
```

Пример 2: Источник для аминокислотной последовательности:

```
<INSDSeq_feature-table>
  <INSDFeature>
    <INSDFeature_key>SOURCE</INSDFeature_key>
    <INSDFeature_location>1..174</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>
</INSDSeq_feature-table>
```

80. Если последовательность является природной и исходный организм принадлежит к известному роду на латыни, но вид организма не указан или не опознан, то в качестве значения квалификатора должно указываться латинское обозначение рода, сопровождаемое “sp.”.

Например:

```
<INSDQualifier_name>organism</INSDQualifier_name>  
<INSDQualifier_value>Bacillus sp.</INSDQualifier_value>
```

81. Если последовательность является природной и исходный организм не имеет обозначения рода и вида на латыни, то значение квалификатора организма должно указываться как “unidentified” («неопознанный»), сопровождаемый любой таксонометрической информацией в квалификаторе “note” для нуклеотидных последовательностей и квалификаторе “NOTE” для аминокислотных последовательностей. Например:

```
<INSDQualifier_name>organism</INSDQualifier_name>  
<INSDQualifier_value>unidentified</INSDQualifier_value>  
<INSDQualifier_name>note</INSDQualifier_name>  
<INSDQualifier_value>bacterium B8</INSDQualifier_value>
```

82. Если последовательность является природной и исходный организм не имеет обозначения рода и вида на латыни, такой как вирус, то должно быть использовано другое приемлемое научное название (например “Canine adenovirus type 2”) как значение квалификатора. Например:

```
<INSDQualifier_name>organism</INSDQualifier_name>  
<INSDQualifier_value>Canine adenovirus type 2</INSDQualifier_value>
```

83. Если последовательность не природная, значение квалификатора организма должно быть указано как “synthetic construct”. Дальнейшая информация в отношении способа генерации последовательности может быть указана с помощью квалификатора “note” для нуклеотидной последовательности и квалификатора “NOTE” для аминокислотной последовательности. Например:

```
<INSDSeq_feature-table>  
  <INSDFeature>  
    <INSDFeature_key>SOURCE</INSDFeature_key>  
    <INSDFeature_location>1..40</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>  
        <INSDQualifier_name>NOTE</INSDQualifier_name>  
        <INSDQualifier_value>synthetic peptide used as assay for  
          antibodies</INSDQualifier_value>  
      </INSDQualifier>  
    </INSDFeature_qual>  
  </INSDFeature>  
</INSDSeq_feature-table>
```

84. Квалификатор “mol_type” для нуклеотидных последовательностей (см. Приложение I, Глава 6) и “MOL_TYPE” для аминокислотных последовательностей (см. Приложение I, Глава 8) должен раскрывать тип молекулы, представленной в последовательности. Данный квалификатор отличается от элемента `INSDSeq_moltype`, описанного в параграфе 54:

(a) Для нуклеотидной последовательности значением квалификатора “mol_type” должно быть одно из перечисленных: “genomic DNA”, “genomic RNA”, “mRNA”, “tRNA”, “rRNA”, “other RNA”, “other DNA”, “transcribed RNA”, “viral cRNA”, “unassigned DNA”, или “unassigned RNA”. Если последовательность неприродного происхождения, например, значением квалификатора “organism” является “synthetic construct”, значением квалификатора “mol_type” должно быть или “other RNA” или “other DNA”;

(b) Для аминокислотных последовательностей значением квалификатора “MOL_TYPE” является “protein”.

Свободный текст

85. Свободный текст - это тип формата для некоторых квалификаторов (как указано в Приложении 1), представленный в форме фразы описательного текста, который должен быть предпочтительно на английском языке.

86. Использование свободного текста должно быть ограничено несколькими короткими терминами, необходимыми для понимания характеристик последовательности. Для каждого квалификатора свободный текст не должен превышать 1000 знаков (символов).

Кодирующие последовательности.

87. Для идентификации кодирующих последовательностей может быть использован функциональный ключ “CDS”, например последовательности нуклеотидов, которые соответствуют последовательности аминокислот в белке и стоп кодон. Элемент `INSDFeature_location` должен определять местоположение признака «CDS» и включать стоп кодон.

88. Квалификаторы “transl_table” и “translation” могут быть использованы с функциональным ключом “CDS” (смотри Приложение 1). Если квалификатор “transl_table” не применяется, допускается использование Таблицы стандартных кодов (смотри Приложение 1, глава 9, таблица 5).

89. Квалификатор «transl_except» должен использоваться с функциональным ключом «CDS» и квалификатором “translation” («перевод») для идентификации кодона, который кодирует пирролизин или селеноцистеин.

90. Аминокислотная последовательность, закодированная кодирующей последовательностью и раскрытая в квалификаторе “translation”, охваченном параграфом 7, должна быть включена в перечень последовательностей с присвоенным собственным идентификационным номером. Идентификационный номер последовательности, присвоенный аминокислотной последовательности, должен быть представлен как значение в квалификаторе “protein_id” с функциональным ключом “CDS”. Квалификатор “ORGANISM” функционального ключа “SOURCE” для аминокислотной последовательности должен быть идентичен своей кодирующей последовательности.

Например:

```
<INSDSeq_feature-table>
  <INSDFeature>
    <INSDFeature_key>CDS</INSDFeature_key>
    <INSDFeature_location>1..507</INSDFeature_location>
    <INSDFeature_qual>
      <INSDQualifier>
        <INSDQualifier_name>transl_table</INSDQualifier_name>
        <INSDQualifier_value>11</INSDQualifier_value>
      </INSDQualifier>
      <INSDQualifier>
        <INSDQualifier_name>translation</INSDQualifier_name>
        <INSDQualifier_value>
          MLVHLERTTIMFDFSSLINLPLIWGLLIAIAVLLYILMDGFDLGIGILLPFAPSDKCRDHMISSIAPFWDGNETWLV
          LGGGGLFAAFPLAYSILMPAFYIPIIIMLLGLIVRGVSFEFRFKAEGKYRRLWDYAFHFGSLGA AFCQGMILGAFIH
          GVEVNGRNFSGQLM
        </INSDQualifier_value>
      </INSDQualifier>
      <INSDQualifier>
        <INSDQualifier_name>protein_id</INSDQualifier_name>
        <INSDQualifier_value>89</INSDQualifier_value>
      </INSDQualifier>
    </INSDFeature_qual>
  </INSDFeature>
</INSDSeq_feature-table>
```

Варианты

91. Исходная последовательность и любые варианты этой последовательности, раскрытые перечислением её остатков и охваченные параграфом 7, должны быть включены в перечень последовательностей со своим собственным присвоенным идентификационным номером последовательности.

92. Любой вариант последовательности, раскрытой как одна последовательность с перечислением альтернативных вариантов остатков на одной или нескольких позициях, должен быть включен в перечень последовательностей и должен быть представлен одной последовательностью, в которой перечислены альтернативные варианты остатков, представленных наиболее ограничивающим неоднозначным символом (см. параграфы 15 и 27).

93. Любой вариант последовательности, раскрытой только посредством ссылки на удаление(я), добавление(я) или замещение(я) в исходной последовательности, следует включать в перечень последовательностей. Такие варианты последовательностей, включенные в перечень последовательностей:

- а) могут быть представлены путём аннотирования исходной последовательности, если она содержит вариант(ы) на отдельной позиции или множестве разных позиций и появление этих вариантов происходит независимо;
- б) должны быть представлены как отдельная последовательность с собственным присвоенным идентификационным номером последовательности, если они содержат варианты на множестве разных позиций и возникновение данных вариантов является взаимосвязанным; и

с) должны быть представлены как отдельная последовательность с собственным присвоенным идентификационным номером последовательности, в случае, если в последовательности содержатся замещения и добавления в более чем 1000 остатков. (см. параграф 86).

94. Таблица ниже показывает правильный способ использования функциональных ключей и квалификаторов для вариантов нуклеиновой кислоты и аминокислоты:

Тип последовательности	Функциональный ключ	Квалификатор	Использование
Нуклеиновая кислота	variation	replace or note	Природные мутации и полиморфизмы, например Alleles, RFLPs
Нуклеиновая кислота	misc_difference	replace or note	Вариабельность, вносимая искусственно, т.е. путём генетических манипуляций или путём химического синтеза
Аминокислота	VAR_SEQ	NOTE	Вариант, полученный путем проведения альтернативного сплайсинга, использования альтернативного промотора, альтернативного инициирования и рибосомного сдвига рамки генетического кода
Аминокислота	VARIANT	NOTE	Любой вариант, для которого VAR_SEQ не применим.

95. Аннотирование последовательности для определённого варианта должно включать функциональный ключ и квалификатор, как указано в таблице выше, и местоположение признака. Значением квалификатора "replace" должен быть только один альтернативный нуклеотид или нуклеотидная последовательность, использующая только символы, определённые в Разделе 1, Таблице 1. Перечень альтернативных вариантов остатков может быть представлен как значение в квалификаторах "note" или "NOTE". В частности, перечень альтернативных аминокислот должен быть представлен как значение в квалификаторе "NOTE", где «X» используется в последовательности, но представляет собой подгруппу «любой из 'A', 'R', 'N', 'D', 'C', 'Q', 'E', 'G', 'H', 'I', 'L', 'K', 'M', 'F', 'P', 'O', 'S', 'U', 'T', 'W', 'Y' или 'V' ". Удаление должно быть представлено пустым значением квалификатора для квалификатора "replace" или указанием в "note" или "NOTE" что остаток может быть удалён. Замещение или добавление остатка(ов) должно быть представлено в квалификаторах "replace", "note" или "NOTE". Для квалификаторов "replace", "note" или "NOTE" форматом значений является свободный текст, который не должен превышать 1000 знаков, как представлено в параграфе 86. Смотри параграф 98 для последовательностей, охваченных параграфом 7, которые даются как замещение или добавление значения квалификатора.

96. Символы, определённые в Приложении 1 (см. главы с 1 по 4, таблицы с 1 по 4 соответственно), если необходимо, должны быть использованы для представления вариантов остатков. Если вариант остатка является модифицированным остатком, не определённым в таблицах 2 или 4 Приложения 1, тогда полное несокращённое название модифицированного остатка должно быть представлено в качестве значения квалификатора для квалификаторов “note” или “NOTE”. Модифицированные остатки должны быть в дальнейшем описаны в таблице характеристик, как предусмотрено в параграфах 17 или 30.

97. Следующие примеры иллюстрируют представление вариантов, согласно параграфам 93 – 96 выше.

Пример 1: Функциональный ключ «misc_difference» для перечисленных альтернативных вариантов нуклеотидов. “n” на позиции 53 последовательности может быть одним из пяти альтернативных нуклеотидов.

```
<INSDFeature>
  <INSDFeature_key>misc_difference</INSDFeature_key>
  <INSDFeature_location>53</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>note</INSDQualifier_name>
      <INSDQualifier_value>w, cmnm5s2u, mam5u, mcm5s2u, or
p</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>modified_base</INSDFeature_key>
  <INSDFeature_location>53</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>mod_base</INSDQualifier_name>
      <INSDQualifier_value>OTHER</INSDQualifier_value>
      <INSDQualifier_name>note</INSDQualifier_name>
      <INSDQualifier_value>cmnm5s2u, mam5u, mcm5s2u, or
p</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
```

Пример 2: Функциональный ключ “misc_difference” для удаления в нуклеотидной последовательности. Нуклеотид в положении 413 удалён.

```
<INSDFeature>
  <INSDFeature_key>misc_difference</INSDFeature_key>
  <INSDFeature_location>413</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>replace</INSDQualifier_name>
      <INSDQualifier_value></INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
```

Пример 3: Функциональный ключ “misc_difference” для добавления в нуклеотидной последовательности. Последовательность “atgccaaatat” добавлена между положениями 100 и 101 исходной последовательности.

```
<INSDFeature>
  <INSDFeature_key>misc_difference</INSDFeature_key>
  <INSDFeature_location>100^101</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>replace</INSDQualifier_name>
      <INSDQualifier_value>atgccaaatat</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
```

Пример 4: Функциональный ключ “variation” для замещения в нуклеотидной последовательности. Цитозин заменяет нуклеотид, указанный на позиции 413 последовательности.

```
<INSDFeature>
  <INSDFeature_key>variation</INSDFeature_key>
  <INSDFeature_location>413</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>replace</INSDQualifier_name>
      <INSDQualifier_value>c</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
```

Пример 5: Функциональный ключ “VARIANT” для замещения в аминокислотной последовательности. Аминокислота в положении 100 может быть заменена на I, A, F, Y, aIle, MeIle или Nle.

```
<INSDFeature>
  <INSDFeature_key>VARIANT</INSDFeature_key>
  <INSDFeature_location>100</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>I, A, F, Y, aIle, MeIle, or Nle</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>MOD_RES</INSDFeature_key>
  <INSDFeature_location>100</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>aIle, MeIle, or Nle</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
```

Пример 6: Функциональный ключ “VARIANT” для замещения в аминокислотной последовательности. Аминокислота в положении 100 может быть заменена на любую аминокислоту, кроме Lys, Arg или His.

```
<INSDFeature>
  <INSDFeature_key>VARIANT</INSDFeature_key>
  <INSDFeature_location>100</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>not K, R, or H</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
```

98. Последовательность, охваченная параграфом 7, которая при аннотировании исходной последовательности представлена как дополнение или замещение в значении квалификатора, должна также быть включена в перечень последовательностей с присвоенным собственным идентификационным номером последовательности.

[Приложение 1 к Стандарту ST.26 следует]

Приложение 1

Контролируемая лексика

Версия 1.1

*Одобен Комитетом по Стандартам ВОИС (КСВ)
на своей пятой сессии 2 июня 2017*

Содержание

Глава 1 Перечень нуклеотидов	2
Глава 2 Перечень модифицированных нуклеотидов	2
Глава 3 Перечень аминокислот	5
Глава 4 Перечень модифицированных и нетипичных аминокислот	6
Глава 5 Функциональные ключи для нуклеотидных последовательностей	7
Глава 6 Квалификаторы для нуклеотидных последовательностей	25
Глава 7 Функциональные ключи для аминокислотных последовательностей	50
Глава 8 Квалификаторы для аминокислотных последовательностей	58
Глава 9 Таблицы генетических кодов	59

1 Перечень нуклеотидов

Коды основных нуклеотидов, которые используются в перечне последовательностей, представлены в Таблице 1. Символ “t” будет истолкован как тимин в ДНК и урацил в РНК, если он используется без последующего описания. В случае, когда присвоен неоднозначный символ (представляющий два и более альтернативных основания), должен быть использован наиболее ограничивающий символ. Например, в случае если основаниями данной позиции могут быть “a или g”, следует использовать скорее “r”, чем “n”. Символ “n” будет рассматриваться как “a или c или g или t/u”, если он используется без дополнительного описания.

Таблица 1: Перечень нуклеотидов

Символ	Нуклеотид
a	аденин /adenine
c	цитозин /cytosine
g	гуанин /guanine
t	тимин в ДНК/урацил в РНК (t/u)thymine in DNA/uracil in RNA (t/u)
m	a или c
r	a или g
w	a или t/u
s	c или g
y	c или t/u
k	g или t/u
v	a или c или g; не t/u
h	a или c или t/u; не g
d	a или g или t/u; не c
b	c или g или t/u; не a
n	a или c или g или t/u; неизвестная или другая

2 Перечень модифицированных нуклеотидов

Аббревиатуры, перечисленные в Таблице 2, являются единственными разрешёнными значениями для квалификатора mod_base. Если определённый модифицированный нуклеотид не представлен в нижеприведённой таблице, тогда в качестве его значения должна использоваться аббревиатура “OTHER”. Если аббревиатурой является “OTHER”, то полное несокращённое название модифицированного основания должно быть приведено в квалификаторе “note”. Аббревиатуры, приведенные в Таблице 2, не должны быть использованы в самой последовательности.

Таблица 2: Перечень модифицированных нуклеотидов

Аббревиатура	Модифицированный нуклеотид
ac4c	4-ацетилцитидин 4-acetylcytidine
chm5u	5-(карбоксигидроксиметил)уридин 5-(carboxyhydroxymethyl)uridine
cm	2'-O-метилцитидин 2'-O-methylcytidine

Аббревиатура	Модифицированный нуклеотид
cmnm5s2u	5-карбоксиметиламинометил-2-тиоуридин (5-carboxymethylaminomethyl-2-thiouridine)
cmnm5u	5-карбоксиметиламинометилуридин (5-carboxymethylaminomethyluridine)
dhu	дигидроуридин (dihydrouridine)
fm	2'-О-метилпсевдоуридин (2'-O-methylpseudouridine)
gal q	бета,D-галактозилквеуозин (beta,D-galactosylqueuosine)
gm	2'-О-метилгуанозин (2'-O-methylguanosine)
i	инозин (inosine)
i6a	N6-изопениладенозин (N6-isopentenyladenosine)
m1a	1-метиладенозин (1-methyladenosine)
m1f	1-метилпсевдоуридин (1-methylpseudouridine)
m1g	1-метилгуанозин (1-methylguanosine)
m1i	1-метиинозин (1-methylinosine)
m22g	2,2-диметилгуанозин (2,2-dimethylguanosine)
m2a	2-метиладенозин (2-methyladenosine)
m2g	2-метилгуанозин (2-methylguanosine)
m3c	3-метилцитидин (3-methylcytidine)
m4c	N4-methylcytosine (N4-methylcytosine)
m5c	5-метилцитидин (5-methylcytidine)
m6a	N6-метиладенозин (N6-methyladenosine)
m7g	7-метилгуанозин (7-methylguanosine)
mam5u	5-метиламинометилуридин (5-methylaminomethyluridine)
mam5s2u	5- метоксиаминометил-2-тиоуридин (5-methoxyaminomethyl-2-thiouridine)
man q	бета,D-манозилквеуозин (beta,D-mannosylqueuosine)

Аббревиатур	Модифицированный нуклеотид
mcm5s2u	5-метиламинометил-2-тиоуридин (5-methylaminomethyl-2-thiouridine)
mcm5u	5-метоксикарбонилметилуридин (5-methoxycarbonylmethyluridine)
mo5u	5-метоксиуридин (5-methoxyuridine)
ms2i6a	2-метилтио-N6-изопентиладенозин (2-methylthio-N6-isopentenyladenosine)
ms2t6a	N-((9-бета-D-рибофуранозил-2- метилтиопурин-6-ил) карбомил)треонин (N-((9-beta-D-ribofuranosyl-2-methylthiopurine-6-yl)carbamoyl)threonine)
mt6a	N-((9-бета-D-рибофураносилпурин-6-ил)N- метилкарбомил)треонин (N-((9-beta-D-ribofuranosylpurine-6-yl)N-methyl-carbamoyl)threonine)
mv	уридин-5-оксоуксусная кислота-метилэфир (uridine-5-oxoacetic acid-methylester)
o5u	уридин-5-оксоуксусная кислота (v) (uridine-5-oxoacetic acid (v))
osyw	вибутоксозин (wybutoxosine)
p	псевдоуридин (pseudouridine)
q	квеуозин (queuosine)
s2c	2-тиоцитидин (2-thiocytidine)
s2t	5-метил-2-тиоуридин (5-methyl-2-thiouridine)
s2u	2-тиоуридин (2-thiouridine_)
s4u	4-тиоуридин (4-thiouridine)
m5u	5-метилуридин (5-methyluridine)
t6a	N-((9-бета-D-рибофураназилпурин-6-ил)карбомил)треонин (N-((9-beta-D-ribofuranosylpurine-6-yl)carbamoyl)threonine)
tm	2'-О-метил-5-метилуридин (2'-O-methyl-5-methyluridine)
um	2'-О-метилуридин (2'-O-methyluridine)
yw	вибутозин (wybutosine)
x	3-(3-амино-3- карбоксипропил)уридин, (аср3)u (3-(3-amino-3-carboxypropyl)uridine, (аср3)u)
OTHER	(требуется квалификатор "note")

3 Перечень аминокислот

Коды аминокислот, которые используются в последовательностях, представлены в Таблице 3. Когда присвоен неоднозначный символ (представляющий две или более альтернативных аминокислоты), должен быть использован наиболее ограничивающий символ. Например, если аминокислота в данной позиции может быть аспарагиновой кислотой или аспарагином, следует использовать скорее символ “B”, чем “X”. Символ “X” будет рассматриваться как один из символов “A”, “R”, “N”, “D”, “C”, “Q”, “E”, “G”, “H”, “I”, “L”, “K”, “M”, “F”, “P”, “O”, “S”, “U”, “T”, “W”, “Y”, или “V”, если он используется без дополнительного описания.

Таблица 3: Перечень аминокислот

Символ	Аминокислота
A	Аланин/Alanine
R	Аргинин/Arginine
N	Аспарагин/Asparagine
D	Аспарагиновая кислота (Аспартат)/ Aspartic acid (Aspartate)
C	Цистеин /Cysteine
Q	Глутамин /Glutamine
E	Глутаминовая кислота (Глутамат)/ Glutamic acid (Glutamate)
G	Глицин/Glycine
H	Гистидин/Histidine
I	Изолейцин/Isoleucine
L	Лейцин/Leucine
K	Лизин/Lysine
M	Метионин/Methionine
F	Фенилаланин/Phenylalanine
P	Пролин/Proline
O	Пирролизин/Pyrrolysine
S	Серин/Serine
U	Селеноцистеин/Selenocysteine
T	Треонин/Threonine
W	Триптофан/Tryptophan
Y	Тирозин/Tyrosine
V	Валин/Valine
B	Аспарагиновая кислота или Аспарагин/Aspartic acid or Asparagine
Z	Глутамин или Глутаминовая кислота/Glutamine or Glutamic acid
J	Лейцин или Изолейцин/Leucine or Isoleucine
X	A или R или N или D или C или Q или E или G или H или I или L или K или M или F или P или O или S или U или T или W или Y или V; “unknown” или “other”

4 Перечень модифицированных и нетипичных аминокислот

В Таблице 4 перечислены единственные разрешённые аббревиатуры для модифицированных или нетипичных (необычных) аминокислот в обязательном квалификаторе “NOTE” для функциональных ключей “MOD_RES” или “SITE”. Значение для квалификатора “NOTE” должно быть или одной из аббревиатур из этой таблицы, где это возможно, или полным, несокращённым названием модифицированной аминокислоты. Аббревиатуры (или полные названия), приведённые в данной таблице, не должны быть использованы в самой последовательности.

Таблица 4: 4 Перечень модифицированных и нетипичных (необычных) аминокислот

Аббревиатура	Модифицированная или нетипичная аминокислота
Aad	2-аминоадипиновая кислота/2-Aminoadipic acid
bAad	3-аминоадипиновая кислота/3-Aminoadipic acid
bAla	бета-аланин, бета-аминопропионовая кислота/ beta-Alanine, beta-Aminopropionic acid
Abu	2-аминобутириновая кислота/2-Aminobutyric acid
4Abu	4-аминобутириновая кислота, пиперидиновая кислота/ 4-Aminobutyric acid, piperidinic acid
Acp	6-аминокапроиновая кислота/6-Aminocaproic acid
Ahe	2-аминогептаноиновая кислота/2-Aminoheptanoic acid
Aib	2-аминоизобутириновая кислота/2-Aminoisobutyric acid
bAib	3-аминоизобутириновая кислота/3-Aminoisobutyric acid
Apm	2-аминопимелиновая кислота/2-Aminopimelic acid
Dbu	2,4-диаминобутириновая кислота /2,4- Diaminobutyric acid
Des	десмозин/Desmosine
Dpm	2,2'-диаминопимелиновая кислота/2,2'- Diaminopimelic acid
Dpr	2,3-диаминопропионовая кислота/2,3- Diaminopropionic acid
EtGly	N-этилглицин/N-Ethylglycine
EtAsn	N-этиласпарагин/N-Ethylasparagine
Hyl	гидроксилизин/Hydroxylysine
aHyl	алло-гидроксилизин/allo-Hydroxylysine
3Hyp	3-гидроксипролин/3-Hydroxyproline
4Hyp	4-гидроксипролин/4-Hydroxyproline
Ide	Издесмозин/Isodesmosine
aIle	алло-изолейцин/allo-Isoleucine
MeGly	N-метилглицин, саркозин/N-Methylglycine, sarcosine
MeIle	N- метилизолейцин/N-Methylisoleucine
MeLys	6-N-метиллизин/6- N-Methyllysine
MeVal	N-метилвалин/N-Methylvaline
Nva	Норвалин/Norvaline
Nle	Норлейцин/Norleucine
Orn	Орнитин/Ornithine

5 Функциональные ключи для нуклеотидных последовательностей

Данная глава содержит перечень разрешённых функциональных ключей для использования в нуклеотидных последовательностях, и перечень обязательных и необязательных квалификаторов. Функциональные ключи перечислены в алфавитном порядке. Функциональные ключи могут быть использованы для ДНК или для РНК, если другое не указано в поле “Molecule scope”. Некоторые функциональные ключи могут подходить для использования с не природными последовательностями в дополнение к определенным “organism scope”.

Названия функциональных ключей в случае XML последовательности должны использоваться в точности так, как они приводятся в описаниях ниже в поле “Feature key”, за исключением функциональных ключей 3'UTR и 5'UTR. Для функциональных ключей 3'UTR и 5'UTR см. поле “Comment” в описании.

5.1. Feature key	C_region
Definition	constant region of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; includes one or more exons depending on the particular chain
Optional qualifiers	allele gene gene_synonym map note product pseudo pseudogene standard_name
Organism scope	eukaryotes
5.2. Feature key	CDS
Definition	coding sequence; sequence of nucleotides that corresponds with the sequence of amino acids in a protein (location includes stop codon); feature may include amino acid conceptual translation
Optional qualifiers	allele codon_start EC_number exception function gene gene_synonym map note number operon product protein_id pseudo pseudogene ribosomal_slippage

	standard_name translation transl_except transl_table trans_splicing
Comment	codon_start qualifier has valid value of 1 or 2 or 3, indicating the offset at which the first complete codon of a coding feature can be found, relative to the first base of that feature; transl_table defines the genetic code table used if other than the Standard or universal genetic code table; genetic code exceptions outside the range of the specified tables are reported in transl_except qualifier; only one of the qualifiers translation, pseudogene or pseudo are permitted with a CDS feature key; when the translation qualifier is used, the protein_id qualifier is mandatory if the translation product contains four or more specifically defined amino acids

5.3. Feature Key	centromere
Definition	region of biological interest identified as a centromere and which has been experimentally characterized
Optional qualifiers	note standard_name
Comment	the centromere feature describes the interval of DNA that corresponds to a region where chromatids are held and a kinetochore is formed

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

5.5. Feature Key	D_segment
Definition	Diversity segment of immunoglobulin heavy chain, and T-cell receptor beta chain
Optional qualifiers	allele gene gene_synonym map

	note
	product
	pseudo
	pseudogene
	standard_name

Organism scope	eukaryotes
----------------	------------

5.6. Feature Key exon

Definition	region of genome that codes for portion of spliced mRNA, rRNA and tRNA; may contain 5'UTR, all CDSs and 3' UTR
------------	--

Optional qualifiers	allele
	EC_number
	function
	gene
	gene_synonym
	map
	note
	number
	product
	pseudo
	pseudogene
	standard_name
	trans_splicing

5.7. Feature Key gene

Definition	region of biological interest identified as a gene and for which a name has been assigned
------------	---

Optional qualifiers	allele
	function
	gene
	gene_synonym
	map
	note
	operon
	product
	pseudo
	pseudogene
	phenotype
	standard_name
	trans_splicing

Comment	the gene feature describes the interval of DNA that corresponds to a genetic trait or phenotype; the feature is, by definition, not strictly bound to its positions at the ends; it is meant to represent a region where the gene is located.
---------	---

5.8. Feature Key iDNA

Definition	intervening DNA; DNA which is eliminated through any of several
------------	---

	kinds of recombination
Optional qualifiers	allele function gene gene_synonym map note number standard_name
Molecule scope	DNA
Comment	e.g., in the somatic processing of immunoglobulin genes.
5.9. Feature Key	intron
Definition	a segment of DNA that is transcribed, but removed from within the transcript by splicing together the sequences (exons) on either side of it
Optional qualifiers	allele function gene gene_synonym map note number pseudo pseudogene standard_name trans_splicing
5.10. Feature Key	J_segment
Definition	joining segment of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains
Optional qualifiers	allele gene gene_synonym map note product pseudo pseudogene standard_name
Parent Key	CDS
Organism scope	eukaryotes
5.11. Feature Key	mat_peptide
Definition	mature peptide or protein coding sequence; coding sequence for the

	mature or final peptide or protein product following post-translational modification; the location does not include the stop codon (unlike the corresponding CDS)
Optional qualifiers	allele EC_number function gene gene_synonym map note product pseudo pseudogene standard_name
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_classqualifier with the value "ribosome_binding_site" must be used for describing ribosome binding sites
5.13. Feature Key	misc_difference
Definition	featured sequence differs from the presented sequence at this location and cannot be described by any other Difference key (variation, or modified_base)
Optional qualifiers	allele clone compare gene gene_synonym map note phenotype replace standard_name
Comment	the misc_difference feature key must be used to describe variability introduced artificially, e.g. by genetic manipulation or by chemical

synthesis; use the replace qualifier to annotate a deletion, insertion, or substitution. The variation feature key must be used to describe naturally occurring genetic variability.

5.14. Feature Key	misc_feature
Definition	region of biological interest which cannot be described by any other feature key; a new or rare feature
Optional qualifiers	allele function gene gene_synonym map note number phenotype product pseudo pseudogene standard_name
Comment	this key should not be used when the need is merely to mark a region in order to comment on it or to use it in another feature's location

5.15. Feature Key	misc_recomb
Definition	site of any generalized, site-specific or replicative recombination event where there is a breakage and reunion of duplex DNA that cannot be described by other recombination keys or qualifiers of source key (proviral)
Optional qualifiers	allele gene gene_synonym map note recombination_class standard_name
Molecule scope	DNA

5.16. Feature Key	misc_RNA
Definition	any transcript or RNA product that cannot be defined by other RNA keys (prim_transcript, precursor_RNA, mRNA, 5'UTR, 3'UTR, exon, CDS, sig_peptide, transit_peptide, mat_peptide, intron, polyA_site, ncRNA, rRNA and tRNA)
Optional qualifiers	allele function gene gene_synonym map note

	- operon - product - pseudo - pseudogene - standard_name - trans_splicing
5.17. Feature Key	misc_structure
Definition	any secondary or tertiary nucleotide structure or conformation that cannot be described by other Structure keys (stem_loop and D-loop)
Optional qualifiers	- allele - function - gene - gene_synonym - map - note - standard_name
5.18. Feature Key	mobile_element
Definition	region of genome containing mobile elements
Mandatory qualifiers	mobile_element_type
Optional qualifiers	- allele - function - gene - gene_synonym - map - note - rpt_family - rpt_type - standard_name
5.19. Feature Key	modified_base
Definition	the indicated nucleotide is a modified nucleotide and should be substituted for by the indicated molecule (given in the mod_base qualifier value)
Mandatory qualifiers	mod_base
Optional qualifiers	- allele - frequency - gene - gene_synonym - map - note
Comment	value for the mandatory mod_base qualifier is limited to the restricted vocabulary for modified base abbreviations in Section 2 of this Annex.

5.20. Feature Key	mRNA
Definition	messenger RNA; includes 5' untranslated region (5'UTR), coding sequences (CDS, exon) and 3' untranslated region (3'UTR)
Optional qualifiers	allele function gene gene_synonym map note operon product pseudo pseudogene standard_name trans_splicing
5.21. Feature Key	ncRNA
Definition	a non-protein-coding gene, other than ribosomal RNA and transfer RNA, the functional molecule of which is the RNA transcript
Mandatory qualifiers	ncRNA_class
Optional qualifiers	allele function gene gene_synonym map note operon product pseudo pseudogene standard_name trans_splicing
Comment	the ncRNA feature must not be used for ribosomal and transfer RNA annotation, for which the rRNA and tRNA feature keys must be used, respectively
5.22. Feature Key	N_region
Definition	extra nucleotides inserted between rearranged immunoglobulin segments
Optional qualifiers	allele gene gene_synonym map note product pseudo pseudogene

standard_name	
5.23. Feature Key	operon
Definition	region containing polycistronic transcript including a cluster of genes that are under the control of the same regulatory sequences/promoter and in the same biological pathway
Mandatory qualifiers	operon
Optional qualifiers	allele function map note phenotype pseudo pseudogene standard_name
5.24. Feature Key	oriT
Definition	origin of transfer; region of a DNA molecule where transfer is initiated during the process of conjugation or mobilization
Optional qualifiers	allele bound_moiety direction gene gene_synonym map note rpt_family rpt_type rpt_unit_range rpt_unit_seq standard_name
Molecule Scope	DNA
Comment	rep_origin must be used to describe origins of replication; direction qualifier has legal values left, right and both, however only left and right are valid when used in conjunction with the oriT feature; origins of transfer can be present in the chromosome; plasmids can contain multiple origins of transfer
5.25. Feature Key	polyA_site
Definition	site on an RNA transcript to which will be added adenine residues by post-transcriptional polyadenylation
Optional qualifiers	allele gene gene_synonym map note

Organism scope eukaryotes and eukaryotic viruses

5.26. Feature Key precursor_RNA

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

Optional qualifiers allele
function
gene
gene_synonym
map
note
operon
product
standard_name
trans_splicing

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

5.27. Feature Key prim_transcript

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

Optional qualifiers allele
operon
standard_name

5.28. Feature Key primer_bind

Definition non-covalent primer binding site for initiation of replication, transcription, or reverse transcription; includes site(s) for

synthetic e.g., PCR primer elements

Optional qualifiers allele
gene
gene_synonym
map
note
standard_name

Comment used to annotate the site on a given sequence to which a primer molecule binds - not intended to represent the sequence of the primer molecule itself; since PCR reactions most often involve pairs of primers, a single primer_bind key may use the order(location,location) operator with two locations, or a pair of primer_bind keys may be used

5.29. Feature Key	propeptide
Definition	propeptide coding sequence; coding sequence for the domain of a proprotein that is cleaved to form the mature protein product.
	function gene gene_synonym map note product pseudo pseudogene standard_name
5.30. Feature Key	protein_bind
Definition	non-covalent protein binding site on nucleic acid
Mandatory qualifiers	bound_moiety
Optional qualifiers	allele function gene gene_synonym map note operon standard_name
Comment	note that the regulatory feature key and regulatory_class qualifier with the value "ribosome_binding_site" must be used to describe ribosome binding sites
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 pseudo pseudogene standard_name
5.32. Feature Key	repeat_region
Definition	region of genome containing repeating units
Optional qualifiers	allele function gene gene_synonym map note rpt_family

	rpt_type rpt_unit_range rpt_unit_seq satellite standard_name
5.33. Feature Key	rep_origin
Definition	origin of replication; starting site for duplication of nucleic acid to give two identical copies
Optional Qualifiers	allele direction standard_name
Comment	direction qualifier has valid values: left, right, or both
5.34. Feature Key	rRNA
Definition	mature ribosomal RNA; RNA component of the ribonucleoprotein particle (ribosome) which assembles amino acids into proteins
Optional qualifiers	allele function gene gene_synonym map note operon product pseudo pseudogene standard_name
Comment	rRNA sizes should be annotated with the product qualifier
5.35. Feature Key	S_region
Definition	switch region of immunoglobulin heavy chains; involved in the rearrangement of heavy chain DNA leading to the expression of a different immunoglobulin class from the same B-cell
Optional qualifiers	allele gene gene_synonym map note product pseudo pseudogene standard_name

5.36. Feature Key	sig_peptide
Definition	signal peptide coding sequence; coding sequence for an N-terminal domain of a secreted protein; this domain is involved in attaching nascent polypeptide to the membrane leader sequence
Optional qualifiers	allele function gene gene_synonym map note product pseudo pseudogene standard_name
5.37. Feature Key	source
Definition	identifies the source of the sequence; this key is mandatory; every sequence will have a single source key spanning the entire sequence
Mandatory qualifiers	organism mol_type
	Optional qualifiers cell_line cell_type chromosome clone clone_lib collected_by collection_date cultivar dev_stage ecotype environmental_sample germline haplogroup haplotype host identified_by isolate isolation_source lab_host lat_lon macronuclear map mating_type note organelle PCR_primers plasmid pop_variant proviral rearranged

segment
serotype
serovar
sex
strain
sub_clone
sub_species
sub_strain
tissue_lib
tissue_type
variety

Molecule scope any

5.38. Feature Key stem_loop

Definition hairpin; a double-helical region formed by base-pairing between adjacent (inverted) complementary sequences in a single strand of RNA or DNA

Optional qualifiers allele
function
gene
gene_synonym
map
note
operon
standard_name

5.39. Feature Key STS

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

Optional qualifiers allele
gene
gene_synonym
map
note
standard_name

Molecule scope DNA

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

5.40. Feature Key telomere

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

Optional qualifiers note
rpt_type
rpt_unit_range

	rpt_unit_seq standard_name
Comment	the telomere feature describes the interval of DNA that corresponds to a specific structure at the end of the linear eukaryotic chromosome which is required for the integrity and maintenance of the end; this region is unique compared to the rest of the chromosome and represents the physical end of the chromosome
5.41. Feature Key	tmRNA
Definition	transfer messenger RNA; tmRNA acts as a tRNA first, and then as an mRNA that encodes a peptide tag; the ribosome translates this mRNA region of tmRNA and attaches the encoded peptide tag to the C-terminus of the unfinished protein; this attached tag targets the protein for destruction or proteolysis
Optional qualifiers	allele function gene gene_synonym map note product pseudo pseudogene standard_name tag_peptide
5.42. Feature Key	transit_peptide
Definition	transit peptide coding sequence; coding sequence for an N-terminal domain of a nuclear-encoded organellar protein; this domain is involved in post-translational import of the protein into the organelle
Optional qualifiers	allele function gene gene_synonym map note product pseudo pseudogene standard_name
5.43. Feature Key	tRNA
Definition	mature transfer RNA, a small RNA molecule (75-85 bases long) that mediates the translation of a nucleic acid sequence into an amino acid sequence
Optional qualifiers	allele anticodon

	- function - gene - gene_synonym - map - note - product - pseudo - pseudogene - standard_name - trans_splicing
5.44. Feature Key	unsure
Definition	a small region of sequenced bases, generally 10 or fewer in its length, which could not be confidently identified. Such a region might contain called bases (a, t, g, or c), or a mixture of called-bases and uncalled-bases ('n').
Optional qualifiers	- allele - compare - gene - gene_synonym - map - note - replace
Comment	use the replace qualifier to annotate a deletion, insertion, or substitution.
5.45. Feature Key	V_region
Definition	variable region of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; codes for the variable amino terminal portion; can be composed of V_segments, D_segments, N_regions, and J_segments
Optional qualifiers	- allele - gene - gene_synonym - map - note - product - pseudo - pseudogene - standard_name
Organism scope	eukaryotes
5.46. Feature Key	V_segment
Definition	variable segment of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; codes for most of the variable region (V_region) and the last few amino acids of the

	leader peptide
Optional qualifiers	allele gene gene_synonym map note product pseudo pseudogene standard_name
5.47. Feature Key	variation
Definition	a related strain contains stable mutations from the same gene (e.g., RFLPs, polymorphisms, etc.) which differ from the presented sequence at this location (and possibly others)
Optional qualifiers	allele compare frequency gene gene_synonym map note phenotype product replace standard_name
Comment	used to describe alleles, RFLP's, and other naturally occurring mutations and polymorphisms; use the replace qualifier to annotate a deletion, insertion, or substitution; variability arising as a result of genetic manipulation (e.g. site directed mutagenesis) must be described with the misc_difference feature;
5.48. Feature Key	3'UTR
Definition	1) region at the 3' end of a mature transcript (following the stop codon) that is not translated into a protein; 2) region at the 3' end of an RNA virus (following the last stop codon) that is not translated into a protein;
Optional qualifiers	allele function gene gene_synonym map note standard_name trans_splicing
Comment	The apostrophe character has special meaning in XML, and must be substituted with "'" in the value of an element. Thus "3'UTR" must be represented as "3'UTR" in the XML file, i.e., <INSDFeature_key>3'UTR</INSDFeature_key>.

5.49. Feature Key	5'UTR
Definition	1) region at the 5' end of a mature transcript (preceding the initiation codon) that is not translated into a protein; 2) region at the 5' end of an RNA virus (preceding the first initiation codon) that is not translated into a protein;
Optional qualifiers	allele function gene gene_synonym map note standard_name trans_splicing
Comment	The apostrophe character has special meaning in XML, and must be substituted with “&apos;” in the value of an element. Thus “5'UTR” must be represented as “5&apos;UTR” in the XML file, i.e., <INSDFeature_key>5&apos;UTR</INSDFeature_key>.

6 Квалификаторы для нуклеотидных последовательностей

Данная глава содержит перечень квалификаторов, которые будут использоваться для описания признаков в нуклеотидных последовательностях. Квалификаторы перечислены в алфавитном порядке.

Если в описании квалификатора (например, germline) в поле Value format указано “none”, элемент INSDQualifier_value не должен использоваться.

ПОЖАЛУЙСТА, ОБРАТИТЕ ВНИМАНИЕ: Любое значение квалификатора, предусмотренное для квалификатора со значением формата “free text” («свободный текст»), может потребовать перевода для национальных / региональных процедур.

6.1. Qualifier	allele
Definition	name of the allele for the given gene
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>adh1-1</INSDQualifier_value>
Comment	all gene-related features (exon, CDS etc) for a given gene should share the same allele qualifier value; the allele qualifier value must, by definition, be different from the gene qualifier value; when used with the variation feature key, the allele qualifier value should be that of the variant.
6.2. Qualifier	anticodon
Definition	location of the anticodon of tRNA and the amino acid for which it codes
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>
6.3. Qualifier	bound_moiety
Definition	name of the molecule/complex that may bind to the given feature
Value format	free text (NOTE: this value may require translation for National/Regional

	procedures)
Example	<INSDQualifier_value>GAL4</INSDQualifier_value>
Comment	A single bound_moiety qualifier is legal on the "misc_binding", "oriT" and "protein_bind" features.
6.4. Qualifier	cell_line
Definition	cell line from which the sequence was obtained
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>MCF7</INSDQualifier_value>
6.5. Qualifier	cell_type
Definition	cell type from which the sequence was obtained
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>leukocyte</INSDQualifier_value>
6.6. Qualifier	chromosome
Definition	chromosome (e.g. Chromosome number) from which the sequence was obtained
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>1</INSDQualifier_value> <INSDQualifier_value>X</INSDQualifier_value>
6.7. Qualifier	clone
Definition	clone from which the sequence was obtained
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>lambda-hIL7.3</INSDQualifier_value>
Comment	a source feature must not contain more than one clone should qualifier; where the sequence was obtained from multiple clones it may be further described in the feature table using the feature key misc_feature and a note qualifier to specify the multiple clones.

6.8. ualifier	clone_lib
Definition	clone library from which the sequence was obtained
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>lambda-hIL7</INSDQualifier_value>
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.
Value format	1 or 2 or 3
Example	<INSDQualifier_value>2</INSDQualifier_value>
6.10. Qualifier	collected_by
Definition	name of persons or institute who collected the specimen
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>Dan Janzen</INSDQualifier_value>
6.11. Qualifier	collection_date
Definition	date that the specimen was collected
Value format	YYYY-MM-DD, YYYY-MM or YYYY
Example	<INSDQualifier_value>1952-10-21</INSDQualifier_value> <INSDQualifier_value>1952-10</INSDQualifier_value> <INSDQualifier_value>1952</INSDQualifier_value>
Comment	'YYYY' is a four-digit value representing the year. 'MM' is a two-digit value representing the month. 'DD' is a two-digit value representing the day of the month.
6.12. Qualifier	compare
Definition	Reference details of an existing public INSD entry to which a comparison is made
Value format	[accession-number.sequence-version]
Example	<INSDQualifier_value>AJ634337.1</INSDQualifier_value>

Comment	This qualifier may be used on the following features: misc_difference, unsure, and variation. Multiple compare qualifiers with different contents are allowed within a single feature. This qualifier is not intended for large-scale annotation of variations, such as SNPs.
---------	---

6.13. Qualifier

cultivar

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

Value format	free text (NOTE: this value may require translation for National/Regional procedures)
--------------	--

Example	<pre><INSDQualifier_value>Nipponbare</INSDQualifier_value> <INSDQualifier_value>Tenuifolius</INSDQualifier_value> <INSDQualifier_value>Candy Cane</INSDQualifier_value> <INSDQualifier_value>IR36</INSDQualifier_value></pre>
---------	---

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

6.14. Qualifier

dev_stage

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

Value format	free text (NOTE: this value may require translation for National/Regional procedures)
--------------	--

Example	<pre><INSDQualifier_value>fourth instar larva</INSDQualifier_value></pre>
---------	---

6.15. Qualifier

direction

Definition	direction of DNA replication
------------	------------------------------

Value format	left, right, or both where left indicates toward the 5' end of the sequence (as presented) and right indicates toward the 3' end
--------------	---

Example	<pre><INSDQualifier_value>left</INSDQualifier_value></pre>
---------	--

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

6.16. Qualifier

EC_number

Definition	Enzyme Commission number for enzyme product of sequence
------------	---

Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>1.1.2.4</INSDQualifier_value> <INSDQualifier_value>1.1.2.-</INSDQualifier_value> <INSDQualifier_value>1.1.2.n</INSDQualifier_value>
Comment	valid values for EC numbers are defined in the list prepared by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) (published in Enzyme Nomenclature 1992, Academic Press, San Diego, or a more recent revision thereof). The format represents a string of four numbers separated by full stops; up to three numbers starting from the end of the string may be replaced by dash "-" to indicate uncertain assignment. Symbol "n" may be used in the last position instead of a number where the EC number is awaiting assignment. Please note that such incomplete EC numbers are not approved by NC-IUBMB.

6.17. Qualifier	ecotype
Definition	a population within a given species displaying genetically based, phenotypic traits that reflect adaptation to a local habitat
Value Format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>Columbia</INSDQualifier_value>
Comment	an example of such a population is one that has adapted hairier than normal leaves as a response to an especially sunny habitat. 'Ecotype' is often applied to standard genetic stocks of Arabidopsis thaliana, but it can be applied to any sessile organism.

6.18. Qualifier	environmental_sample
Definition	identifies sequences derived by direct molecular isolation from a bulk environmental DNA sample (by PCR with or without subsequent cloning of the product, DGGE, or other anonymous methods) with no reliable identification of the source organism. Environmental samples include clinical samples, gut contents, and other sequences from anonymous organisms that may be associated with a particular host. They do not include endosymbionts that can be reliably recovered from a particular host, organisms from a readily identifiable but uncultured field sample (e.g., many cyanobacteria), or phytoplasmas that can be reliably recovered from diseased plants (even though these cannot be grown in axenic culture)
Value format	none
Comment	used only with the source feature key; source feature keys containing the environmental_sample qualifier should also contain the isolation_source qualifier; a source feature including the

environmental_sample qualifier must not include the strain qualifier.

6.19. Qualifier	exception
Definition	indicates that the coding region cannot be translated using standard biological rules
Value format	One of the following controlled vocabulary phrases: RNA editing rearrangement required for product annotated by transcript or proteomic data
Example	<INSDQualifier_value>RNA editing</INSDQualifier_value> <INSDQualifier_value>rearrangement required for product</INSDQualifier_value>
Comment	only to be used to describe biological mechanisms such as RNA editing; protein translation of a CDS with an exception qualifier will be different from the corresponding conceptual translation; must not be used where transl_except qualifier would be adequate, e.g. in case of stop codon completion use.
6.20. Qualifier	frequency
Definition	frequency of the occurrence of a feature
Value format	free text representing the proportion of a population carrying the feature expressed as a fraction (NOTE: this value may require translation for National/Regional procedures)
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
Value format	free text (NOTE: 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.
6.22. Qualifier	gene
Definition	symbol of the gene corresponding to a sequence region

Value format	free text (NOTE: this value may require translation for National/Regional procedures)
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.

6.23. Qualifier	gene_synonym
Definition	synonymous, replaced, obsolete or former gene symbol
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>Hox-3.3</INSDQualifier_value> in a feature where the gene qualifier value is Hoxc6
Comment	used where it is helpful to indicate a gene symbol synonym; when the gene_synonym qualifier is used, a primary gene symbol must always be indicated in a gene qualifier

6.24. Qualifier	germline
Definition	the sequence presented has not undergone somatic rearrangement as part of an adaptive immune response; it is the unrearranged sequence that was inherited from the parental germline
Value format	none
Comment	germline qualifier must not be used to indicate that the source of the sequence is a gamete or germ cell; germline and rearranged qualifiers must not be used in the same source feature; germline and rearranged qualifiers must only be used for molecules that can undergo somatic rearrangements as part of an adaptive immune response; these are the T-cell receptor (TCR) and immunoglobulin loci in the jawed vertebrates, and the unrelated variable lymphocyte receptor (VLR) locus in the jawless fish (lampreys and hagfish); germline and rearranged qualifiers should not be used outside of the Craniata (taxid=89593)

6.25. Qualifier	haplogroup
Definition	name for a group of similar haplotypes that share some sequence variation. Haplogroups are often used to track migration of population groups.
Value format	free text (NOTE: this value may require translation for National/Regional procedures)

Example	<INSDQualifier_value>H*</INSDQualifier_value>
6.26. Qualifier	haplotype
Definition	name for a specific set of alleles that are linked together on the same physical chromosome. In the absence of recombination, each haplotype is inherited as a unit, and may be used to track gene flow in populations.
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>Dw3 B5 Cw1 A1</INSDQualifier_value>
6.27. Qualifier	host
Definition	natural (as opposed to laboratory) host to the organism from which sequenced molecule was obtained
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>Homo sapiens</INSDQualifier_value> <INSDQualifier_value>Homo sapiens 12 year old girl</INSDQualifier_value> <INSDQualifier_value>Rhizobium NGR234</INSDQualifier_value>
6.28. Qualifier	identified_by
Definition	name of the expert who identified the specimen taxonomically
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>John Burns</INSDQualifier_value>
6.29. Qualifier	isolate
Definition	individual isolate from which the sequence was obtained
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>Patient #152</INSDQualifier_value> <INSDQualifier_value>DGGE band PSBAC-13</INSDQualifier_value>
6.30. Qualifier	isolation_source
Definition	describes the physical, environmental and/or local geographical

	source of the biological sample from which the sequence was derived
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Examples	<INSDQualifier_value>rumen isolates from standard Pelleted ration-fed steer #67</INSDQualifier_value> <INSDQualifier_value>permanent Antarctic sea ice</INSDQualifier_value> <INSDQualifier_value>denitrifying activated sludge from carbon_limited continuous reactor</INSDQualifier_value>
Comment	used only with the source feature key; source feature keys containing an environmental_sample qualifier should also contain an isolation_source qualifier

6.31. Qualifier	lab_host
Definition	scientific name of the laboratory host used to propagate the source organism from which the sequenced molecule was obtained
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>Gallus gallus</INSDQualifier_value> <INSDQualifier_value>Gallus gallus embryo</INSDQualifier_value> <INSDQualifier_value>Escherichia coli strain DH5 alpha</INSDQualifier_value> <INSDQualifier_value>Homo sapiens HeLa cells</INSDQualifier_value>
Comment	the full binomial scientific name of the host organism should be used when known; extra conditional information relating to the host may also be included

6.32. Qualifier	lat_lon
Definition	geographical coordinates of the location where the specimen was collected
Value format	free text - degrees latitude and longitude in format "d[d.ddd] N S d[dd.ddd] W E" (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>47.94 N 28.12 W</INSDQualifier_value> <INSDQualifier_value>45.0123 S 4.1234 E</INSDQualifier_value>

6.33. Qualifier	macronuclear
Definition	if the sequence shown is DNA and from an organism which undergoes chromosomal differentiation between macronuclear and micronuclear stages, this qualifier is used to denote that the sequence is from macronuclear DNA

value format none

6.34. Qualifier

map

Definition genomic map position of feature

Value format free text
(NOTE: this value may require translation for National/Regional procedures)

Example <INSDQualifier_value>8q12- q13</INSDQualifier_value>

6.35. Qualifier

mating_type

Definition mating type of the organism from which the sequence was obtained; mating type is used for prokaryotes, and for eukaryotes that undergo meiosis without sexually dimorphic gametes

Value format free text
(NOTE: this value may require translation for National/Regional procedures)

Examples
<INSDQualifier_value>MAT-1</INSDQualifier_value>
<INSDQualifier_value>plus</INSDQualifier_value>
<INSDQualifier_value>-</INSDQualifier_value>
<INSDQualifier_value>odd</INSDQualifier_value>
<INSDQualifier_value>even</INSDQualifier_value>

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

6.36. Qualifier

mobile_element_type

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

Value format <mobile_element_type>[:<mobile_element_name>]
where <mobile_element_type> is one of the following:
transposon
retrotransposon
integron
insertion sequence
non-LTR retrotransposon
SINE
MITE
LINE
other

Example <INSDQualifier_value>transposon:Tnp9</INSDQualifier_value>

Comment mobile_element_type is legal on mobile_element feature key only. Mobile element should be used to represent both elements which are currently mobile, and those which were mobile in the past. value

"other" for <mobile_element_type> requires a <mobile_element_name>

6.37. Qualifier	mod_base
Definition	abbreviation for a modified nucleotide base
Value format	modified base abbreviation chosen from this Annex, Section 2
Example	<INSDQualifier_value>m5c</INSDQualifier_value> <INSDQualifier_value>OTHER</INSDQualifier_value>
Comment	specific modified nucleotides not found in paragraph 2 of this Annex are annotated by entering OTHER as the value for the mod_base qualifier and including a note qualifier with the full name of the modified base as its value
6.38. Qualifier	mol_type
Definition	molecule type of sequence
Value format	One chosen from the following: genomic DNA genomic RNA mRNA tRNA rRNA other RNA other DNA transcribed RNA viral cRNA unassigned DNA unassigned RNA
Example	<INSDQualifier_value>genomic DNA</INSDQualifier_value> <INSDQualifier_value>other RNA</INSDQualifier_value>
Comment	mol_type qualifier is mandatory on the source feature key; the value "genomic DNA" does not imply that the molecule is nuclear (e.g. organelle and plasmid DNA must be described using "genomic DNA"); ribosomal RNA genes must be described using "genomic DNA"; "rRNA" must only be used if the ribosomal RNA molecule itself has been sequenced; values "other RNA" and "other DNA" must be applied to synthetic molecules, values "unassigned DNA", "unassigned RNA" must be applied where in vivo molecule is unknown.
6.39. Qualifier	ncRNA_class
Definition	a structured description of the classification of the non-coding RNA described by the ncRNA parent key
Value format	TYPE where TYPE is one of the following controlled vocabulary terms or phrases: antisense_RNA autocatalytically_spliced_intron

	ribozyme hammerhead_ribozyme RNase_P_RNA RNase_MRP_RNA telomerase_RNA guide_RNA rasiRNA sCRNA siRNA miRNA piRNA snoRNA snRNA SRP_RNA" vault_RNA Y_RNA other
Example	<pre><INSDQualifier_value>autocatalytically_spliced_intron </INSDQualifier_value> <INSDQualifier_value>siRNA</INSDQualifier_value> <INSDQualifier_value>sCRNA</INSDQualifier_value> <INSDQualifier_value>other</INSDQualifier_value></pre>
Comment	specific ncRNA types not yet in the ncRNA_class controlled vocabulary must be annotated by entering other as the ncRNA_class qualifier value, and providing a brief explanation of novel ncRNA_class in a note qualifier
6.40. Qualifier	note
Definition	any comment or additional information
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<pre><INSDQualifier_value>A comment about the feature</INSDQualifier_value></pre>
6.41. Qualifier	number
Definition	a number to indicate the order of genetic elements (e.g. exons or introns) in the 5' to 3' direction
Value format	free text (with no whitespace characters) (NOTE: this value may require translation for National/Regional procedures)
Example	<pre><INSDQualifier_value>4</INSDQualifier_value> <INSDQualifier_value>6B</INSDQualifier_value></pre>
Comment	text limited to integers, letters or combination of integers and/or letters represented as a data value that contains no whitespace characters; any additional terms should be included in a

standard_name qualifier. Example: a number qualifier with a value of 2A and a standard_name qualifier with a value of long

6.42. Qualifier	operon
Definition	name of the group of contiguous genes transcribed into a single transcript to which that feature belongs
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>lac</INSDQualifier_value>
6.43. Qualifier	organelle
Definition	type of membrane-bound intracellular structure from which the sequence was obtained
Value format	One of the following controlled vocabulary terms and phrases: chromatophore hydrogenosome mitochondrion nucleomorph plastid mitochondrion:kinetoplast plastid:chloroplast plastid:apicoplast plastid:chromoplast plastid:cyanelle plastid:leucoplast plastid:proplastid,
Examples	<INSDQualifier_value>chromatophore</INSDQualifier_value> <INSDQualifier_value>hydrogenosome</INSDQualifier_value> <INSDQualifier_value>mitochondrion</INSDQualifier_value> <INSDQualifier_value>nucleomorph</INSDQualifier_value> <INSDQualifier_value>plastid</INSDQualifier_value> <INSDQualifier_value>mitochondrion:kinetoplast</INSDQualifier_value> > <INSDQualifier_value>plastid:chloroplast</INSDQualifier_value> <INSDQualifier_value>plastid:apicoplast</INSDQualifier_value> <INSDQualifier_value>plastid:chromoplast</INSDQualifier_value> <INSDQualifier_value>plastid:cyanelle</INSDQualifier_value> <INSDQualifier_value>plastid:leucoplast</INSDQualifier_value> <INSDQualifier_value>plastid:proplastid</INSDQualifier_value>
6.44. Qualifier	organism
Definition	scientific name of the organism that provided the sequenced genetic material, if known, or the available taxonomic information if the organism is unclassified; or an indication that the sequence is a synthetic construct
Value format	free text

	(NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>Homo sapiens</INSDQualifier_value>
6.45. Qualifier	PCR_primers
Definition	PCR primers that were used to amplify the sequence. A single /PCR_primers qualifier should contain all the primers used for a single PCR reaction. If multiple forward or reverse primers are present in a single PCR reaction, multiple sets of fwd_name/fwd_seq or rev_name/rev_seq values will be present
Value format	[fwd_name: XXX1,]fwd_seq: xxxxx1,[fwd_name: XXX2,]fwd_seq: xxxxx2, [rev_name: YYY1,]rev_seq: yyyyy1,[rev_name: YYY2,]rev_seq: yyyyy2
Example	<INSDQualifier_value>fwd_name: CO1P1, fwd_seq: ttgatTTTTGgtcayccwgaagt,rev_name: CO1R4, rev_seq: ccwvytardcctarraartgttg</INSDQualifier_value> <INSDQualifier_value>fwd_name: hoge1, fwd_seq: cgkgtgtatcttact, rev_name: hoge2, rev_seq: cg<lt;i>gt;gtgtatcttact</INSDQualifier_value> <INSDQualifier_value>fwd_name: CO1P1, fwd_seq: ttgatTTTTGgtcayccwgaagt, fwd_name: CO1P2, fwd_seq: gatacacaggtcayccwgaagt, rev_name: CO1R4, rev_seq: ccwvytardcctarraartgttg</INSDQualifier_value>
Comment	fwd_seq and rev_seq are both mandatory; fwd_name and rev_name are both optional.Both sequences must be presented in 5'>3' order. The sequences must be given in the symbols from Annex 1, paragraph 1, except for the modified bases, which must be enclosed within angle brackets < >. In XML, the angle brackets < and > must be substituted with < and > since they are reserved characters in XML.
6.46. Qualifier	phenotype
Definition	phenotype conferred by the feature, where phenotype is defined as a physical, biochemical or behavioural characteristic or set of characteristics
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>erythromycin resistance</INSDQualifier_value>
6.47. Qualifier	plasmid
Definition	name of naturally occurring plasmid from which the sequence was obtained, where plasmid is defined as an independently replicating genetic unit that cannot be described by chromosome or segment qualifiers

Value format	free text (NOTE: this value may require translation for National/Regional procedures)
--------------	--

Example	<INSDQualifier_value>pc589</INSDQualifier_value>
---------	--

6.48. Qualifier	pop_variant
-----------------	-------------

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

Value format	free text (NOTE: this value may require translation for National/Regional procedures)
--------------	--

Example	<INSDQualifier_value>pop1</INSDQualifier_value> <INSDQualifier_value>Bear Paw</INSDQualifier_value>
---------	--

6.49. Qualifier	product
-----------------	---------

Definition	name of the product associated with the feature, e.g. the mRNA of an mRNA feature, the polypeptide of a CDS, the mature peptide of a mat_peptide, etc.
------------	--

Value format	free text (NOTE: this value may require translation for National/Regional procedures)
--------------	--

Example	<INSDQualifier_value>trypsinogen</INSDQualifier_value> (when qualifier appears in CDS feature) <INSDQualifier_value>trypsin</INSDQualifier_value> (when qualifier appears in mat_peptide feature) <INSDQualifier_value>XYZ neural-specific transcript</INSDQualifier_value> (when qualifier appears in mRNA feature)
---------	--

6.50. Qualifier	protein_id
-----------------	------------

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

Value format	an integer greater than zero
--------------	------------------------------

Example	<INSDQualifier_value>89</INSDQualifier_value>
---------	---

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
--------------	------

6.52. Qualifier	pseudo
Definition	indicates that this feature is a non-functional version of the element named by the feature key
Value format	none
Comment	The qualifier pseudo should be used to describe non-functional genes that are not formally described as pseudogenes, e.g. CDS has no translation due to other reasons than pseudogenization events. Other reasons may include sequencing or assembly errors. In order to annotate pseudogenes the qualifier pseudogene must be used, indicating the TYPE of pseudogene.
6.53. Qualifier	pseudogene
Definition	indicates that this feature is a pseudogene of the element named by the feature key
Value format	TYPE where TYPE is one of the following controlled vocabulary terms or phrases: processed unprocessed unitary allelic unknown
Example	<pre><INSDQualifier_value>processed</INSDQualifier_value> <INSDQualifier_value>unprocessed</INSDQualifier_value> <INSDQualifier_value>unitary</INSDQualifier_value> <INSDQualifier_value>allelic</INSDQualifier_value> <INSDQualifier_value>unknown</INSDQualifier_value></pre>
Comment	Definitions of TYPE values: processed - the pseudogene has arisen by reverse transcription of a mRNA into cDNA, followed by reintegration into the genome. Therefore, it has lost any intron/exon structure, and it might have a pseudo-polyA-tail. unprocessed - the pseudogene has arisen from a copy of the parent gene by duplication followed by accumulation of random mutations. The changes, compared to their functional homolog, include insertions, deletions, premature stop codons, frameshifts and a higher proportion of non-synonymous versus synonymous substitutions. unitary - the pseudogene has no parent. It is the original gene, which is functional in some species but disrupted in some way (indels, mutation, recombination) in another species or strain. allelic - a (unitary) pseudogene that is stable in the population but importantly it has a functional alternative allele also in the population. i.e., one strain may have the gene, another strain may have the pseudogene. MHC haplotypes have allelic pseudogenes. unknown - the submitter does not know the method of pseudogenization.

6.54. Qualifier	rearranged
Definition	the sequence presented in the entry has undergone somatic rearrangement as part of an adaptive immune response; it is not the unrearranged sequence that was inherited from the parental germline
Value format	none
Comment	The rearranged qualifier must not be used to annotate chromosome rearrangements that are not involved in an adaptive immune response; germline and rearranged qualifiers must not be used in the same source feature; germline and rearranged qualifiers must only be used for molecules that can undergo somatic rearrangements as part of an adaptive immune response; these are the T-cell receptor (TCR) and immunoglobulin loci in the jawed vertebrates, and the unrelated variable lymphocyte receptor (VLR) locus in the jawless fish (lampreys and hagfish); germline and rearranged qualifiers should not be used outside of the Craniata (taxid=89593)
6.55. Qualifier	recombination_class
Definition	a structured description of the classification of recombination hotspot region within a sequence
Value format	TYPE where TYPE is one of the following controlled vocabulary terms or phrases: mitotic_recombination non_allelic_homologous_recombination_region chromosome_breakpoint
Example	<INSDQualifier_value>meiotic recombination</INSDQualifier_value> <INSDQualifier_value>chromosome_breakpoint</INSDQualifier_value>
Comment	specific recombination classes not yet in the recombination_class controlled vocabulary must be annotated by entering "other" as the recombination_class qualifier value and providing a brief explanation of the novel recombination_class in a note qualifier
6.56. Qualifier	regulatory_class
Definition	a structured description of the classification of transcriptional, translational, replicational and chromatin structure related regulatory elements in a sequence
Value format	TYPE where TYPE is one of the following controlled vocabulary terms or phrases: DNase_I_hypersensitive_site

	enhancer_blocking_element imprinting_control_region insulator locus_control_region matrix_attachment_region minus_35_signal minus_10_signal recoding_stimulatory_region replication_regulatory_region response_element polyA_signal_sequence ribosome_binding_site riboswitch silencer TATA_box transcriptional_cis_regulatory_region other
Example	<INSDQualifier_value>promoter</INSDQualifier_value> <INSDQualifier_value>enhancer</INSDQualifier_value> <INSDQualifier_value>ribosome_binding_site</INSDQualifier_value>
Comment	specific regulatory classes not yet in the regulatory_class controlled vocabulary must be annotated by entering “other” as the regulatory_class qualifier value and providing a brief explanation of the novel regulatory_class in a note qualifier
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
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>a</INSDQualifier_value> <INSDQualifier_value></INSDQualifier_value> - for a deletion
6.58. Qualifier	ribosomal_slippage
Definition	during protein translation, certain sequences can program ribosomes to change to an alternative reading frame by a mechanism known as ribosomal slippage
Value format	none
Comment	a join operator, e.g.: [join(486..1784,1787..4810)] must be used in the CDS feature location to indicate the location of ribosomal_slippage

6.59. Qualifier	rpt_family
Definition	type of repeated sequence; "Alu" or "Kpn", for example
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>Alu</INSDQualifier_value>
6.60. Qualifier	rpt_type
Definition	structure and distribution of repeated sequence
Value format	One of the following controlled vocabulary terms or phrases: tandem direct inverted flanking nested dispersed long_terminal_repeat non_ltr_retrotransposon_polymeric_tract centromeric_repeat telomeric_repeat x_element_combinatorial_repeat y_prime_element other
Example	<INSDQualifier_value>inverted</INSDQualifier_value> <INSDQualifier_value>long_terminal_repeat</INSDQualifier_value>
Comment	Definitions of the values: tandem - a repeat that exists adjacent to another in the same orientation; direct - a repeat that exists not always adjacent but is in the same orientation; inverted - a repeat pair occurring in reverse orientation to one another on the same molecule; flanking - a repeat lying outside the sequence for which it has functional significance (eg. transposon insertion target sites); nested - a repeat that is disrupted by the insertion of another element; dispersed - a repeat that is found dispersed throughout the genome; terminal - a repeat at the ends of and within the sequence for which it has functional significance (eg. transposon LTRs); long_terminal_repeat - a sequence directly repeated at both ends of a defined sequence, of the sort typically found in retroviruses; non_ltr_retrotransposon_polymeric_tract - a polymeric tract, such as poly(dA), within a non LTR retrotransposon; centromeric_repeat - a repeat region found within the modular centromere; telomeric_repeat - a repeat region found within the telomere; x_element_combinatorial_repeat - a repeat region located between

the X element and the telomere or adjacent Y' element;
y_prime_element - a repeat region located adjacent to telomeric repeats or X element combinatorial repeats, either as a single copy or tandem repeat of two to four copies;

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

6.61. Qualifier	rpt_unit_range
Definition	location of a repeating unit expressed as a range
Value format	<base_range> - where <base_range> is the first and last base (separated by two dots) of a repeating unit
Example	<INSDQualifier_value>202..245</INSDQualifier_value>
Comment	used to indicate the base range of the sequence that constitutes a repeating unit within the region specified by the feature keys oriT and repeat_region.
6.62. Qualifier	rpt_unit_seq
Definition	identity of a repeat sequence
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>aagggc</INSDQualifier_value> <INSDQualifier_value>ag(5)tg(8)</INSDQualifier_value> <INSDQualifier_value>(AAAGA)6(AAAA)1(AAAGA)12</INSDQualifier_value>
Comment	used to indicate the literal sequence that constitutes a repeating unit within the region specified by the feature keys oriT and repeat_region
6.63. Qualifier	satellite
Definition	identifier for a satellite DNA marker, compose of many tandem repeats (identical or related) of a short basic repeated unit
Value format	<satellite_type>[:<class>][<identifier>] - where <satellite_type> is one of the following: satellite; microsatellite; minisatellite
Example	<INSDQualifier_value>satellite: Slα</INSDQualifier_value> <INSDQualifier_value>satellite: α</INSDQualifier_value> <INSDQualifier_value>satellite: γ III</INSDQualifier_value> <INSDQualifier_value>microsatellite: DC130</INSDQualifier_value>
Comment	many satellites have base composition or other properties that differ from those of the rest of the genome that allows them to be

identified.

6.64. Qualifier	segment
Definition	name of viral or phage segment sequenced
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>6</INSDQualifier_value>
6.65. Qualifier	serotype
Definition	serological variety of a species characterized by its antigenic properties
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>B1</INSDQualifier_value>
Comment	used only with the source feature key; the Bacteriological Code recommends the use of the term 'serovar' instead of 'serotype' for the prokaryotes; see the International Code of Nomenclature of Bacteria (1990 Revision) Appendix 10.B "Infraspecific Terms".
6.66. Qualifier	serovar
Definition	serological variety of a species (usually a prokaryote) characterized by its antigenic properties
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>O157:H7</INSDQualifier_value>
Comment	used only with the source feature key; the Bacteriological Code recommends the use of the term 'serovar' instead of 'serotype' for prokaryotes; see the International Code of Nomenclature of Bacteria (1990 Revision) Appendix 10.B "Infraspecific Terms".
6.67. Qualifier	sex
Definition	sex of the organism from which the sequence was obtained; sex is used for eukaryotic organisms that undergo meiosis and have sexually dimorphic gametes
Value format	free text (NOTE: this value may require translation for National/Regional procedures)

Examples	<code><INSDQualifier_value>female</INSDQualifier_value></code> <code><INSDQualifier_value>male</INSDQualifier_value></code> <code><INSDQualifier_value>hermaphrodite</INSDQualifier_value></code> <code><INSDQualifier_value>unisexual</INSDQualifier_value></code> <code><INSDQualifier_value>bisexual</INSDQualifier_value></code> <code><INSDQualifier_value>asexual</INSDQualifier_value></code> <code><INSDQualifier_value>monoecious</INSDQualifier_value></code> [or monecious] <code><INSDQualifier_value>dioecious</INSDQualifier_value></code> [or diecious]
Comment	The sex qualifier should be used (instead of mating_type qualifier) in the Metazoa, Embryophyta, Rhodophyta & Phaeophyceae; mating_type qualifier should be used (instead of sex qualifier) in the Bacteria, Archaea & Fungi; neither sex nor mating_type qualifiers should be used in the viruses; outside of the taxa listed above, mating_type qualifier should be used unless the value of the qualifier is taken from the vocabulary given in the examples above

6.68. Qualifier	standard_name
Definition	accepted standard name for this feature
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<code><INSDQualifier_value>dotted</INSDQualifier_value></code>
Comment	use standard_name qualifier to give full gene name, but use gene qualifier to give gene symbol (in the above example gene qualifier value is Dt).

6.69. Qualifier	strain
Definition	strain from which sequence was obtained
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<code><INSDQualifier_value>BALB/c</INSDQualifier_value></code>
Comment	feature entries including a strain qualifier must not include the environmental_sample qualifier

6.70. Qualifier	sub_clone
Definition	sub-clone from which sequence was obtained
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<code><INSDQualifier_value>lambda-hIL7.20g</INSDQualifier_value></code>

Comment	a source feature must not contain more than one sub_clone qualifier; to indicate that the sequence was obtained from multiple sub_clones, multiple sources may be further described using the feature key "misc_feature" and the qualifier "note"
---------	---

6.71. Qualifier	sub_species
-----------------	-------------

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

Value format	free text (NOTE: this value may require translation for National/Regional procedures)
--------------	--

Example	<INSDQualifier_value>lactis</INSDQualifier_value>
---------	---

6.72. Qualifier	sub_strain
-----------------	------------

Definition	name or identifier of a genetically or otherwise modified strain from which sequence was obtained, derived from a parental strain (which should be annotated in the strain qualifier). sub_strain from which sequence was obtained
------------	--

Value format	free text (NOTE: this value may require translation for National/Regional procedures)
--------------	--

Example	<INSDQualifier_value>abis</INSDQualifier_value>
---------	---

Comment	must be accompanied by a strain qualifier in a source feature; if the parental strain is not given, the modified strain should be annotated in the strain qualifier instead of sub_strain. For example, either a strain qualifier with the value K-12 and a substrain qualifier with the value MG1655 or a strain qualifier with the value MG1655
---------	---

6.73. Qualifier	tag_peptide
-----------------	-------------

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

Value format	<base_range> - where <base_range> provides the first and last base (separated by two dots) of the location for the proteolysis tag
--------------	--

Example	<INSDQualifier_value>90..122</INSDQualifier_value>
---------	--

Comment	it is recommended that the amino acid sequence corresponding to the tag_peptide be annotated by describing a 5' partial CDS feature; e.g. CDS with a location of <90..122
---------	---

6.74. Qualifier	tissue_lib
-----------------	------------

Definition	tissue library from which sequence was obtained
------------	---

Value format	free text
--------------	-----------

	(NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>tissue library 772</INSDQualifier_value>
6.75. Qualifier	tissue_type
Definition	tissue type from which the sequence was obtained
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>liver</INSDQualifier_value>
6.76. Qualifier	transl_except
Definition	translational exception: single codon the translation of which does not conform to genetic code defined by organism or transl_table.
Value format	(pos:location,aa:<amino_acid>) where <amino_acid> is the three letter abbreviation for the amino acid coded by the codon at the base_range position
Example	<INSDQualifier_value>(pos:213..215,aa:Trp) </INSDQualifier_value> <INSDQualifier_value>(pos:462..464,aa:OTHER) </INSDQualifier_value> <INSDQualifier_value>(pos:1017,aa:TERM) </INSDQualifier_value> <INSDQualifier_value>(pos:2000..2001,aa:TERM) </INSDQualifier_value>
Comment	if the amino acid is not one of the specific amino acids listed in Annex 1, paragraph 3, use OTHER as <amino_acid> and provide the name of the unusual amino acid in a note qualifier; for modified amino-acid selenocysteine use three letter abbreviation 'Sec' (one letter symbol 'U' in amino-acid sequence) for <amino_acid>; for modified amino-acid pyrrolysine use three letter abbreviation 'Pyl' (one letter symbol 'O' in amino-acid sequence) for <amino_acid>; for partial termination codons where TAA stop codon is completed by the addition of 3' A residues to the mRNA either a single base_position or a base_range is used for the location, see the third and fourth examples above, in conjunction with a note qualifier indicating 'stop codon completed by the addition of 3' A residues to the mRNA'.
6.77. Qualifier	transl_table
Definition	definition of genetic code table used if other than universal or standard genetic code table. Tables used are described in this Annex
Value format	<integer> where <integer> is the number assigned to the genetic code table
Example	<INSDQualifier_value>3</INSDQualifier_value> - example where the yeast mitochondrial code is to be used

Comment if the transl_table qualifier is not used to further annotate a CDS feature key, then the CDS is translated using the Standard Code (i.e. Universal Genetic Code). Genetic code exceptions outside the range of specified tables are reported in transl_except qualifiers.

6.78. Qualifier trans_splicing

Definition indicates that exons from two RNA molecules are ligated in intermolecular reaction to form mature RNA

Value format none

Comment should be used on features such as CDS, mRNA and other features that are produced as a result of a trans-splicing event. This qualifier must be used only when the splice event is indicated in the "join" operator, e.g. join(complement(69611..69724),139856..140087) in the feature location

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

Value format contiguous string of one-letter amino acid abbreviations from this Annex paragraph 3, "x" is to be used for AA exceptions.

Example <INSDQualifier_value>MASTFPWPYRGCASTPSLKGLIMCTW</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.

6.80. Qualifier variety

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

Value format free text
(NOTE: this value may require translation for National/Regional procedures)

Example <INSDQualifier_value>insularis</INSDQualifier_value>

Comment use the cultivar qualifier for cultivated plant varieties, i.e., products of artificial selection; varieties other than plant and fungal variatas should be annotated via a note qualifier, e.g. with the value <INSDQualifier_value>breed:Cukorova</INSDQualifier_value>

7 Функциональные ключи для аминокислотных последовательностей

Данная глава содержит перечень возможных функциональных ключей, которые будут использоваться в аминокислотных последовательностях. Функциональные ключи перечислены в алфавитном порядке.

7.1. Feature Key	ACT_SITE
Definition	Amino acid(s) involved in the activity of an enzyme
Optional qualifiers	NOTE
Comment	Each amino acid residue of the active site must be annotated separately with the ACT_SITE feature key. The corresponding amino acid residue number must be provided as the location descriptor in the feature location element.
7.2. Feature Key	BINDING
Definition	Binding site for any chemical group (co-enzyme, prosthetic group, etc.). The chemical nature of the group is indicated in the NOTE qualifier
Mandatory qualifiers	NOTE
Comment	Examples of values for the "NOTE" qualifier: "Heme (covalent)" and "Chloride." Where appropriate, the features keys CA_BIND, DNA_BIND, METAL, and NP_BIND should be used rather than BINDING.
7.3. Feature Key	CA_BIND
Definition	Extent of a calcium-binding region
Optional qualifiers	NOTE
7.4. Feature Key	CARBOHYD
Definition	Glycosylation site
Mandatory qualifiers	NOTE
Comment	This key describes the occurrence of the attachment of a glycan (mono- or polysaccharide) to a residue of the protein. The type of linkage (C-, N- or O-linked) to the protein is indicated in the "NOTE" qualifier. If the nature of the reducing terminal sugar is known, its abbreviation is shown between parentheses. If three dots '...' follow the abbreviation this indicates an extension of the carbohydrate chain. Conversely no dots means that a monosaccharide is linked. Examples of values used in the "NOTE" qualifier: N-linked (GlcNAc...); O-linked (GlcNAc); O-linked (Glc...); C-linked (Man)partial; O-linked (Ara...).

7.5. Feature Key	CHAIN
Definition	Extent of a polypeptide chain in the mature protein
Optional qualifiers	NOTE
7.6. Feature Key	COILED
Definition	Extent of a coiled-coil region
Optional qualifiers	NOTE
7.7. Feature Key	COMPBIAS
Definition	Extent of a compositionally biased region
Optional qualifiers	NOTE
7.8. Feature Key	CONFLICT
Definition	Different sources report differing sequences.
Optional qualifiers	NOTE
Comment	Examples of values for the "NOTE" qualifier: Missing; K -> Q; GSDSE -> RIRLR; V -> A.
7.9. Feature Key	CROSSLNK
Definition	Post translationally formed amino acid bonds.
Mandatory qualifiers	NOTE
Comment	Covalent linkages of various types formed between two proteins (interchain cross-links) or between two parts of the same protein (intrachain cross-links); except for cross-links formed by disulfide bonds, for which the "DISULFID" feature key is to be used. For an interchain cross-link, the location descriptor in the feature location element is the residue number of the amino acid cross-linked to the other protein. For an intrachain cross-link, the location descriptors in the feature location element are the residue numbers of the cross-linked amino acids in conjunction with the "join" location operator, e.g. "join(42,50)." The NOTE qualifier indicates the nature of the cross-link; at least specifying the name of the conjugate and the identity of the two amino acids involved. Examples of values for the "NOTE" qualifier: "Isoglutamyl cysteine thioester (Cys-Gln);" "Beta-methylanthionine (Cys-Thr);" and "Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in ubiquitin)"
7.10. Feature Key	DISULFID
Definition	Disulfide bond

Optional qualifiers NOTE

Comment For an interchain disulfide bond, the location descriptor in the feature location element is the residue number of the cysteine linked to the other protein. For an intrachain cross-link, the location descriptors in the feature location element are the residue numbers of the linked cysteines in conjunction with the “join” location operator, e.g. “join(42,50)”. For interchain disulfide bonds, the NOTE qualifier indicates the nature of the cross-link, by identifying the other protein, for example, “Interchain (between A and B chains)”

7.11. Feature Key DNA_BIND

Definition Extent of a DNA-binding region

Mandatory qualifiers NOTE

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

7.12. Feature Key DOMAIN

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

Mandatory qualifiers NOTE

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

7.13. Feature Key HELIX

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

Optional qualifiers NOTE

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

7.14. Feature Key INIT_MET

Definition Initiator methionine

Optional qualifiers NOTE

Comment The location descriptor in the feature location element is “1”. This

feature key indicates the N-terminal methionine is cleaved off. This feature is not used when the initiator methionine is not cleaved off.

7.15. Feature Key	INTRAMEM
Definition	Extent of a region located in a membrane without crossing it
Optional qualifiers	NOTE
7.16. Feature Key	LIPID
Definition	Covalent binding of a lipid moiety
Mandatory qualifiers	NOTE
Comment	The chemical nature of the bound lipid moiety is given in the NOTE qualifier, indicating at least the name of the lipidated amino acid. Examples of values for the "NOTE" qualifier: "N-myristoyl glycine"; "GPI-anchor amidated serine" and "S-diacylglycerol cysteine."
7.17. Feature Key	METAL
Definition	Binding site for a metal ion.
Mandatory qualifiers	NOTE
Comment	The NOTE qualifier indicates the nature of the metal. Examples of values for the "NOTE" qualifier: "Iron (heme axial ligand)" and "Copper".
7.18. Feature Key	MOD_RES
Definition	Posttranslational modification of a residue
Mandatory qualifiers	NOTE
Comment	The chemical nature of the modified residue is given in the NOTE qualifier, indicating at least the name of the post-translationally modified amino acid. If the modified amino acid is listed in Section 4 of this Annex, the abbreviation may be used in place of the full name. Examples of values for the "NOTE" qualifier: "N-acetylalanine"; "3-Hyp"; and "MeLys" or "N-6-methyllysine"
7.19. Feature Key	MOTIF
Definition	Short (up to 20 amino acids) sequence motif of biological interest
Optional qualifiers	NOTE
7.20. Feature Key	MUTAGEN
Definition	Site which has been experimentally altered by mutagenesis

Optional qualifiers	NOTE
7.21. Feature Key	NON_STD
Definition	Non-standard amino acid
Optional qualifiers	NOTE
Comment	This key describes the occurrence of non-standard amino acids selenocysteine (U) and pyrrolysine (O) in the amino acid sequence.
7.22. Feature Key	NON_TER
Definition	The residue at an extremity of the sequence is not the terminal residue
Optional qualifiers	NOTE
Comment	If applied to position 1, this means that the first position is not the N-terminus of the complete molecule. If applied to the last position, it means that this position is not the C-terminus of the complete molecule.
7.23. Feature Key	NP_BIND
Definition	Extent of a nucleotide phosphate-binding region
Mandatory qualifiers	NOTE
Comment	The nature of the nucleotide phosphate is indicated in the NOTE qualifier. Examples of values for the "NOTE" qualifier: "ATP" and "FAD".
7.24. Feature Key	PEPTIDE
Definition	Extent of a released active peptide
Optional qualifiers	NOTE
7.25. Feature Key	PROPEP
Definition	Extent of a propeptide
Optional qualifiers	NOTE
7.26. Feature Key	REGION
Definition	Extent of a region of interest in the sequence
Optional qualifiers	NOTE

7.27. Feature Key	REPEAT
Definition	Extent of an internal sequence repetition
Optional qualifiers	NOTE
7.28. Feature Key	SIGNAL
Definition	Extent of a signal sequence (prepeptide)
Optional qualifiers	NOTE
7.29. Feature Key	SITE
Definition	Any interesting single amino-acid site on the sequence that is not defined by another feature key. It can also apply to an amino acid bond which is represented by the positions of the two flanking amino acids
Mandatory qualifier	NOTE
Comment	When SITE is used to annotate a modified amino acid the value for the qualifier "NOTE" must either be an abbreviation set forth in in Section 4 of this Annex or the complete, unabbreviated name of the modified amino acid.
7.30. Feature Key	SOURCE
Definition	Identifies the source of the sequence; this key is mandatory; every sequence will have a single SOURCE feature spanning the entire sequence
Mandatory qualifiers	MOL_TYPE
	ORGANISM
Optional qualifiers	NOTE
7.31. Feature Key	STRAND
Definition	Secondary structure: Beta-strand; for example Hydrogen bonded beta-strand or residue in an isolated beta-bridge
Optional qualifiers	NOTE
Comment	This feature is used only for proteins whose tertiary structure is known. Only three types of secondary structure are specified: helices (key HELIX), beta-strands (key STRAND) and turns (key TURN). Residues not specified in one of these classes are in a 'loop' or 'random-coil' structure.

7.32. Feature Key	TOPO_DOM
Definition	Topological domain
Optional qualifiers	NOTE
7.33. Feature Key	TRANSMEM
Definition	Extent of a transmembrane region
Optional qualifiers	NOTE
7.34. Feature Key	TRANSIT
Definition	Extent of a transit peptide (mitochondrion, chloroplast, thylakoid, cyanelle, peroxisome etc.)
Optional qualifiers	NOTE
7.35. Feature Key	TURN
Definition	Secondary structure Turns, for example, H-bonded turn (3-turn, 4-turn or 5-turn)
Optional qualifiers	NOTE
Comment	This feature is used only for proteins whose tertiary structure is known. Only three types of secondary structure are specified: helices (key HELIX), beta-strands (key STRAND) and turns (key TURN). Residues not specified in one of these classes are in a 'loop' or 'random-coil' structure.
7.36. Feature Key	UNSURE
Definition	Uncertainties in the sequence
Optional qualifiers	NOTE
Comment	Used to describe region(s) of an amino acid sequence for which the authors are unsure about the sequence presentation.
7.37. Feature Key	VARIANT
Definition	Authors report that sequence variants exist.
Optional qualifiers	NOTE
7.38. Feature Key	VAR_SEQ
Definition	Description of sequence variants produced by alternative splicing, alternative promoter usage, alternative initiation and ribosomal frameshifting

Optional qualifiers NOTE

7.39. Feature Key ZN_FING

Definition Extent of a zinc finger region

Mandatory qualifiers NOTE

Comment The type of zinc finger is indicated in the NOTE qualifier. For example: “GATA-type” and “NR C4-type”

8 Квалификаторы для аминокислотных последовательностей

Данная глава содержит перечень возможных квалификаторов, которые будут использоваться в аминокислотных последовательностях.

ПОЖАЛУЙСТА, ОБРАТИТЕ ВНИМАНИЕ: Любое значение квалификатора, предусмотренное для квалификатора со значением формата “free text” («свободный текст»), может потребовать перевода для национальных / региональных процедур.

8.1. Qualifier	MOL_TYPE
Definition	In vivo molecule type of sequence
Value format	protein
Example	<INSDQualifier_value>protein</INSDQualifier_value>
Comment	The "MOL_TYPE" qualifier is mandatory on the SOURCE feature key.
8.2. Qualifier	NOTE
Definition	Any comment or additional information
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>Heme (covalent)</INSDQualifier_value>
Comment	The “NOTE” qualifier is mandatory for the feature keys: BINDING; CARBOHYD; CROSSLNK; DISULFID; DNA_BIND; DOMAIN; LIPID; METAL; MOD_RES; NP_BIND and ZN_FING
8.3. Qualifier	ORGANISM
Definition	Scientific name of the organism that provided the peptide
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>Homo sapiens</INSDQualifier_value>
Comment	The “ORGANISM” qualifier is mandatory for the SOURCE feature key.

9 Таблицы генетических кодов

Таблица 5 представляет Таблицы Генетических Кодов, которые будут использоваться для преобразования кодирующих последовательностей. Значением для квалификатора `trans_table` является номер, присвоенный соответствующей таблице генетического кода. Где `CDS` описывается квалификатором преобразования, но не квалификатором `transl_table`, для преобразования используется по умолчанию Standard Code (стандартный код) – 1. (Примечание: Таблицы генетических кодов 7, 8, 15 и 17-20 не существуют, поэтому эти числа отсутствуют в Таблице 5).

Table 5: Genetic Code Tables

[illegible]

[illegible]

[illegible]

[illegible]

[Приложение 2 к ST.26 следует]

Приложение 2

ОПРЕДЕЛЕНИЕ ТИПА ДОКУМЕНТА ДЛЯ ПЕРЕЧНЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ (DTD)

Версия 1.1

Одобен Комитетом по Стандартам ВОИС (КСВ)
на своей пятой сессии 2 июня 2017

```
<?xml version="1.0" encoding="UTF-8"?>
<!--Annex II of ST.26, Document Type Definition (DTD) for Sequence Listing

This entity may be identified by the PUBLIC identifier:
*****
PUBLIC "-//WIPO//DTD SEQUENCE LISTING 1.1//EN" "ST26SequenceListing_V1_1.dtd"
*****
* PUBLIC DTD URL

* http://www.wipo.int/standards/DTD/ST26SequenceListing_V1_1.dtd
*****

Revision of Annex II to WIPO Standard ST.26 is submitted for approval by the Committee
on WIPO Standards (CWS) at its fifth session.

*****
* CONTACTS
*****
xml.standards@wipo.int

*****
* NOTES
*****
The sequence data part is a subset of the complete INSDC DTD V.1.5 that only covers
the requirements of WIPO Standard ST.26.

*****
* REVISION HISTORY
*****
2017-06-02: Version 1.1 (if it is approved by the CWS)
Changes:
Comments added to <INSDSeq_length>, <INSDSeq_division> and <INSDSeq_sequence> to
clarify the reason of the differences between the INSDC DTD v.1.5 and ST26 Sequence
Listing DTD V1_1.
*****
2016-03-24: Version 1.0 adopted by the CWS/4Bis
2014-03-11: Final draft for adoption.
*****

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

<!--The elements ApplicantName and InventorName are optional in this DTD to facilitate
the conversion between various encoding schemes-->
<!ATTLIST ST26SequenceListing
```

```
dtdVersion CDATA #REQUIRED
fileName CDATA #IMPLIED
softwareName CDATA #IMPLIED
softwareVersion CDATA #IMPLIED
productionDate CDATA #IMPLIED >

<!--ApplicantFileReference
Applicant's or agent's file reference, mandatory if application identification not
provided.
-->
<!ELEMENT ApplicantFileReference (#PCDATA) >

<!--ApplicationIdentification
Application identification for which the sequence listing is submitted, when
available.
-->
<!ELEMENT ApplicationIdentification (IPOfficeCode?,ApplicationNumberText,
FilingDate?) >

<!--EarliestPriorityApplicationIdentification
Application identification of the earliest claimed priority, which Contains
IPOfficeCode, ApplicationNumberText and FilingDate elements.
-->
<!ELEMENT EarliestPriorityApplicationIdentification (IPOfficeCode?,
ApplicationNumberText,FilingDate?) >

<!--ApplicantName
The name of the first mentioned applicant in characters set forth in paragraph 40 a)
of the ST.26 main body document.
-->
<!--languageCode: Appropriate language code from ISO 639-1 - Codes for the
representation of names of languages - Part 1: Alpha-2
-->
<!ELEMENT ApplicantName (#PCDATA) >
<!ATTLIST ApplicantName
languageCode CDATA #REQUIRED >

<!--ApplicantNameLatin
Where ApplicantName is typed in characters other than those as set forth in paragraph
40 b), a translation or transliteration of the name of the first mentioned applicant
must also be typed in characters as set forth in paragraph 40 b) .
-->
<!ELEMENT ApplicantNameLatin (#PCDATA) >

<!--InventorName
Name of the first mentioned inventor typed in the characters as set forth in paragraph
40 a).-->
<!--languageCode: Appropriate language code from ISO 639-1 - Codes for the
representation of names of languages - Part 1: Alpha-2
-->
<!ELEMENT InventorName (#PCDATA)>
<!ATTLIST InventorName
languageCode CDATA #REQUIRED >

<!--InventorNameLatin
Where InventorName is typed in characters other than those as set forth in paragraph
40 b), a translation or transliteration of the first mentioned inventor may also be
typed in characters as set forth in paragraph 40 b) .
-->
<!ELEMENT InventorNameLatin (#PCDATA) >

<!--InventionTitle
Title of the invention typed in the characters as set forth in paragraph 40 a) in the
language of filing. A translation of the title of the invention into additional
languages may be typed in the characters as set forth in paragraph 40 a) using
additional InventionTitle elements. Preferably two to seven words.
-->
<!--languageCode: Appropriate language code from ISO 639-1 - Codes
for the representation of names of languages - Part 1: Alpha-2
-->
```

```
<!ELEMENT InventionTitle (#PCDATA) >
<!ATTLIST InventionTitle
    languageCode CDATA #REQUIRED >

<!--SequenceTotalQuantity
Indicates the total number of sequences in the document.
Its purpose is to be quickly accessible for automatic processing.
-->
<!ELEMENT SequenceTotalQuantity (#PCDATA) >

<!--SequenceData
Data for individual Sequence.
For intentionally skipped sequences see the ST.26 main body document.
-->
<!ELEMENT SequenceData (INSDSeq) >
<!ATTLIST SequenceData
    sequenceIDNumber CDATA #REQUIRED >

<!--IPOfficeCode
ST.3 code. For example, if the application identification is PCT/IB2013/099999, then
IPOfficeCode value will be IB.
-->
<!ELEMENT IPOfficeCode (#PCDATA) >

<!--ApplicationNumberText
The application identification as provided by the office of filing (eg.
PCT/IB2013/099999)
-->
<!ELEMENT ApplicationNumberText (#PCDATA) >

<!--FilingDate
The date of filing of the patent application for which the sequence listing is
submitted ST.2 format (paragraphs 7 (a) and 11) "CCYY-MM-DD", using a 4-digit calendar
year, a 2-digit calendar month and a 2-digit day within the calendar month, e.g.,
2015-01-31
-->
<!ELEMENT FilingDate (#PCDATA) >

<!--*****
* INSD Part
*****

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

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

This part is based on:

The International Nucleotide Sequence Database (INSD) collaboration.

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

<!--INSDSeq
Sequence data. Changed INSD V1.5 DTD elements, INSDSeq_division and INSDSeq_sequence
from optional to mandatory per business requirements.
-->
<!ELEMENT INSDSeq (INSDSeq_length,INSDSeq_moltype,INSDSeq_division,
    INSDSeq_other-seqids?,INSDSeq_feature-table?,INSDSeq_sequence) >

<!--INSDSeq_length
The length of the sequence.INSDSeq_length allows only integer.
-->
<!ELEMENT INSDSeq_length (#PCDATA) >

<!--INSDSeq_moltype
```

Admissible values: DNA, RNA, AA

-->

<!ELEMENT INSDSeq_moltype (#PCDATA) >

<!--INSDSeq_division

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

-->

<!ELEMENT INSDSeq_division (#PCDATA) >

<!--INSDSeq_other-seqids

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

-->

<!ELEMENT INSDSeq_other-seqids (INSDSeqid?) >

<!--INSDSeq_feature-table

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

-->

<!ELEMENT INSDSeq_feature-table (INSDFeature+) >

<!--INSDSeq_sequence

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

-->

<!ELEMENT INSDSeq_sequence (#PCDATA) >

<!--INSDSeqid

Intended for the use of Patent Offices in data exchange only.

Format:

pat|{office code}|{publication number}|{document kind code}|{Sequence identification number}

where office code is the code of the IP office publishing the patent document, publication number is the publication number of the application or patent, document kind code is the letter codes to distinguish patent documents as defined in ST.16 and Sequence identification number is the number of the sequence in that application or patent

Example:

pat|WO|2013999999|A1|123456

This represents the 123456th sequence from WO patent publication No. 2013999999 (A1)

-->

<!ELEMENT INSDSeqid (#PCDATA) >

<!--INSDFeature

Description of one feature.

-->

<!ELEMENT INSDFeature (INSDFeature_key,INSDFeature_location,INSDFeature_qual?) >

<!--INSDFeature_key

A word or abbreviation indicating a feature.

-->

<!ELEMENT INSDFeature_key (#PCDATA) >

<!--INSDFeature_location

Region of the presented sequence which corresponds to the feature.

-->

<!ELEMENT INSDFeature_location (#PCDATA) >

<!--INSDFeature_qual

List of qualifiers containing auxiliary information about a feature.

-->

```
<!ELEMENT INSDFeature_qual (INSDQualifier*) >

<!--INSDQualifier
Additional information about a feature.
For coding sequences and variants see the ST.26 main body document.
-->
<!ELEMENT INSDQualifier (INSDQualifier_name,INSDQualifier_value?) >

<!--INSDQualifier_name
Name of the qualifier.
-->
<!ELEMENT INSDQualifier_name (#PCDATA) >

<!--INSDQualifier_value
Value of the qualifier.
-->
<!ELEMENT INSDQualifier_value (#PCDATA) >
```

[Приложение 3 следует]

Приложение 3

ОБРАЗЕЦ ПЕРЕЧНЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ (XML файл)

Версия 1.1

Одобен Комитетом по Стандартам ВОИС (КСВ) на своей пятой сессии 2 июня 2017

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ST26SequenceListing PUBLIC "-//WIPO//DTD SEQUENCE LISTING 1.0//EN"
ST26SequenceListing_V1_0.dtd">
<ST26SequenceListing dtdVersion="V1_1" fileName="AnnexIII_Sequence_Listing_Specimen.xml"
softwareName="SEQL-software-name" softwareVersion="1.1" productionDate="2017-06-02">
  <ApplicationIdentification>
    <IPOfficeCode>IB</IPOfficeCode>
    <ApplicationNumberText>PCT/IB2015/099999</ApplicationNumberText>
    <FilingDate>2015-01-31</FilingDate>
  </ApplicationIdentification>
  <ApplicantFileReference>AB123</ApplicantFileReference>
  <EarliestPriorityApplicationIdentification>
    <IPOfficeCode>IB</IPOfficeCode>
    <ApplicationNumberText>PCT/IB2014/111111</ApplicationNumberText>
    <FilingDate>2014-01-30</FilingDate>
  </EarliestPriorityApplicationIdentification>
  <ApplicantName languageCode="ja">出願製薬株式会社</ApplicantName>
  <ApplicantNameLatin>Shutsugan Pharmaceuticals Kabushiki Kaisha</ApplicantNameLatin>
  <InventorName languageCode="ja">特許 太郎</InventorName>
  <InventorNameLatin>Taro Tokkyo</InventorNameLatin>
  <InventionTitle languageCode="ja">efgタンパク質のためのマウスabcd-1遺伝子</InventionTitle>
  <InventionTitle languageCode="en">Mus musculus abcd-1 gene for efg protein</InventionTitle>
  </InventionTitle>
  <SequenceTotalQuantity>11</SequenceTotalQuantity>
  <SequenceData sequenceIDNumber="1">
    <INSDSeq>
      <INSDSeq_length>133</INSDSeq_length>
      <INSDSeq_moltype>DNA</INSDSeq_moltype>
      <INSDSeq_division>PAT</INSDSeq_division>
      <INSDSeq_feature-table>
        <INSDFeature>
          <INSDFeature_key>source</INSDFeature_key>
          <INSDFeature_location>1..133</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>genomic DNA</INSDQualifier_value>
            </INSDQualifier>
          </INSDFeature_qual>
        </INSDFeature>
      </INSDSeq_feature-table>
      <INSDSeq_sequence>
        atgaaattaaacataaaarggatgataaaatgagatttgatataaaaaagggttttagagtttagcagagaaggattttgaga
        cggcatggagagagacaagggcattaataaaggataaacatattgacaata</INSDSeq_sequence>
      </INSDSeq>
    </SequenceData>
    <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>
          <INSDQualifier_name>NOTE</INSDQualifier_name>
          <INSDQualifier_value>Synthetic peptide antigen fragment
          </INSDQualifier_value>
        </INSDQualifier>
      </INSDFeature_qual>
    </INSDFeature>
  </INSDSeq_feature-table>
  <INSDSeq_sequence>GSLSDVRKDVKEKRIDKALEAFKNKMDKEK</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="3">
  <INSDSeq>
    <INSDSeq_length>62</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..62</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>genomic DNA</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>CDS</INSDFeature_key>
        <INSDFeature_location>3..62</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>translation</INSDQualifier_name>
            <INSDQualifier_value>MLAPDCFPDPTRIYSSSLC</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>protein_id</INSDQualifier_name>
            <INSDQualifier_value>4</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
  </INSDSeq>
</SequenceData>

```

```

        <INSDSeq_sequence>tgatgctcgcacctgactgtcccttcgacccccacgcatttatagctccagcctgtgctag
    </INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="4">
    <INSDSeq>
        <INSDSeq_length>19</INSDSeq_length>
        <INSDSeq_moltype>AA</INSDSeq_moltype>
        <INSDSeq_division>PAT</INSDSeq_division>
        <INSDSeq_feature-table>
            <INSDFeature>
                <INSDFeature_key>SOURCE</INSDFeature_key>
                <INSDFeature_location>1..19</INSDFeature_location>
                <INSDFeature_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>
        </INSDSeq_feature-table>
        <INSDSeq_sequence>MLAPDCPFDPTRIYSSSLC</INSDSeq_sequence>
    </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="5">
    <INSDSeq>
        <INSDSeq_length>133</INSDSeq_length>
        <INSDSeq_moltype>DNA</INSDSeq_moltype>
        <INSDSeq_division>PAT</INSDSeq_division>
        <INSDSeq_feature-table>
            <INSDFeature>
                <INSDFeature_key>source</INSDFeature_key>
                <INSDFeature_location>1..133</INSDFeature_location>
                <INSDFeature_qual>
                    <INSDQualifier>
                        <INSDQualifier_name>organism</INSDQualifier_name>
                        <INSDQualifier_value>Solanum lycopersicum</INSDQualifier_value>
                    </INSDQualifier>
                    <INSDQualifier>
                        <INSDQualifier_name>mol_type</INSDQualifier_name>
                        <INSDQualifier_value>genomic DNA</INSDQualifier_value>
                    </INSDQualifier>
                    <INSDQualifier>
                        <INSDQualifier_name>note</INSDQualifier_name>
                        <INSDQualifier_value>common name: tomato</INSDQualifier_value>
                    </INSDQualifier>
                </INSDFeature_qual>
            </INSDFeature>
            <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>
        </INSDSeq_feature-table>
    </INSDSeq>
</SequenceData>

```

```

        <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>
    <INSDFeature>
        <INSDFeature_key>variation</INSDFeature_key>
        <INSDFeature_location>60</INSDFeature_location>
        <INSDFeature_qual>
            <INSDQualifier>
                <INSDQualifier_name>replace</INSDQualifier_name>
                <INSDQualifier_value>c</INSDQualifier_value>
            </INSDQualifier>
        </INSDFeature_qual>
    </INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>
    atgaaattaaaacanaaaaggnatgataaaatgagatttgatataaaaaaggttttagagttagcagagaaggattttgaga
    cggcatggagagagacaagggcattaataaaggataaacatattgacaata</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="6">
    <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>
                        <INSDQualifier_name>NOTE</INSDQualifier_name>
                        <INSDQualifier_value>Synthetic peptide antigen fragment
                        </INSDQualifier_value>
                    </INSDQualifier>
                </INSDFeature_qual>
            </INSDFeature>
            <INSDFeature>
                <INSDFeature_key>MOD_RES</INSDFeature_key>
                <INSDFeature_location>3</INSDFeature_location>
                <INSDFeature_qual>
                    <INSDQualifier>
                        <INSDQualifier_name>NOTE</INSDQualifier_name>
                        <INSDQualifier_value>N-acetylalanine</INSDQualifier_value>
                    </INSDQualifier>
                </INSDFeature_qual>
            </INSDFeature>
        </INSDSeq_feature-table>
    </INSDSeq>
</SequenceData>

```

```

</INSDFeature>
<INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>7</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>Orn</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>13</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>D-Arginine</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>UNSURE</INSDFeature_key>
  <INSDFeature_location>15</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>A or V</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>VARIANT</INSDFeature_key>
  <INSDFeature_location>20</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>I, A, F, Y, aIle, MeIle, or Nle
      </INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>22</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>Homoserine</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>GSASDVXKDVEKRIXKALEXFSNKMDSK</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="7">
  <INSDSeq>
    <INSDSeq_length/>
    <INSDSeq_moltype/>
    <INSDSeq_division/>
    <INSDSeq_sequence>000</INSDSeq_sequence>
  </INSDSeq>

```

```
</SequenceData>
<SequenceData sequenceIDNumber="8">
  <INSDSeq>
    <INSDSeq_length>74</INSDSeq_length>
    <INSDSeq_moltype>RNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..74</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>organism</INSDQualifier_name>
            <INSDQualifier_value>Dengue virus 2</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>mol_type</INSDQualifier_name>
            <INSDQualifier_value>genomic RNA</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>
      atgaaattaaaacataaaaggatgataaaatgagatttgatataaaaaagggttttagagtttagcagagaagga
    </INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="9">
  <INSDSeq>
    <INSDSeq_length>120</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..120</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>organism</INSDQualifier_name>
            <INSDQualifier_value>synthetic construct</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>mol_type</INSDQualifier_name>
            <INSDQualifier_value>other DNA</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>misc_feature</INSDFeature_key>
        <INSDFeature_location>1..60</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>DNA fragment</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>misc_feature</INSDFeature_key>
        <INSDFeature_location>61..120</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>

```

```

        <INSDQualifier_value>RNA fragment</INSDQualifier_value>
      </INSDQualifier>
    </INSDFeature_qual>
  </INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>
cgacccacgctccgaggaaccaaccatcacgtttgaggacttcgtgaaggaattggataatacccgctccctacaaaatgg
cgagcgccgactcattgctctcgtacgcgtcgagcggc</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="10">
  <INSDSeq>
    <INSDSeq_length>288</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..288</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>organism</INSDQualifier_name>
            <INSDQualifier_value>Candida albicans</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>mol_type</INSDQualifier_name>
            <INSDQualifier_value>genomic DNA</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>CDS</INSDFeature_key>
        <INSDFeature_location>1..288</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>translation</INSDQualifier_name>
            <INSDQualifier_value>
MNLTLHNVIQTDSRGEKFMKIPEIYIRGIHIKYLRIIPDDIMGYAKEQSMINMENNRNYQKRRGTSS
GGGGGGGGSGDSRRFNNRQSHGHNYGRR</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>transl_table</INSDQualifier_name>
            <INSDQualifier_value>12</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>protein_id</INSDQualifier_name>
            <INSDQualifier_value>11</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>
atgaatttaacottacataatgtttatacaaacgattcccaggtgagaaatttatgaaaattcccgaaatatatttcgtg
gtatacatattaaatattttaagaattcctgatgatattatgggatatgcaaaagaacaaagtatgataaatatggaaaatg
aatcgatacaaaaaagaagaggtactagcagtggtggtggtggtggtggtggtggtggaagtgtgattcaagaaggttt
aataatagacaactgcattggacataattatggacgtagatga</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="11">
  <INSDSeq>
    <INSDSeq_length>95</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>

```

```
<INSDFeature>
  <INSDFeature_key>SOURCE</INSDFeature_key>
  <INSDFeature_location>1..95</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>ORGANISM</INSDQualifier_name>
      <INSDQualifier_value>Candida albicans</INSDQualifier_value>
    </INSDQualifier>
    <INSDQualifier>
      <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
      <INSDQualifier_value>protein</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>
  MNLTLHNVIQTDSRGEKFMKIPETIYIRGIHIKYLRIIPDDIMGYAKEQSMINMENNRNRYQKRRGTSSGGGGGGGGSGDSRRF
  NNRQSHGHNYGRR</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
</ST26SequenceListing>
```

[Приложение 4 следует]

Приложение 4

Подмножество символов из Таблицы Unicode Basic Latin Code

Версия 1.1

Одобен Комитетом по Стандартам ВОИС (КСВ)
на своей пятой сессии 2 июня 2017

Символ амперсанда (0026) разрешен только как часть заранее определенного значения или как часть ссылки цифрового символа (&#nnnn;). Кавычка (0022), апостроф (0027), знак меньше чем (003C) и знак больше чем (003E) - не разрешены и должны быть представлены их заранее определенными значениями.

Unicode code point	character	name
0020		SPACE
0021	!	EXCLAMATION MARK
0023	#	NUMBER SIGN
0024	\$	DOLLAR SIGN
0025	%	PERCENT SIGN
0026	&	AMPERSAND
0028	(	LEFT PARENTHESIS
0029	)	RIGHT PARENTHESIS
002A	*	ASTERISK
002B	+	PLUS SIGN
002C	,	COMMA
002D	-	HYPHEN-MINUS
002E	.	FULL STOP
002F	/	SOLIDUS
0030	0	DIGIT ZERO
0031	1	DIGIT ONE
0032	2	DIGIT TWO
0033	3	DIGIT THREE
0034	4	DIGIT FOUR
0035	5	DIGIT FIVE
0036	6	DIGIT SIX
0037	7	DIGIT SEVEN
0038	8	DIGIT EIGHT
0039	9	DIGIT NINE
003A	:	COLON

003B	;	SEMICOLON
003D	=	EQUALS SIGN
003F	?	QUESTION MARK
0040	@	COMMERCIAL AT
0041	A	LATIN CAPITAL LETTER A
0042	B	LATIN CAPITAL LETTER B
0043	C	LATIN CAPITAL LETTER C
0044	D	LATIN CAPITAL LETTER D
0045	E	LATIN CAPITAL LETTER E
0046	F	LATIN CAPITAL LETTER F
0047	G	LATIN CAPITAL LETTER G
0048	H	LATIN CAPITAL LETTER H
0049	I	LATIN CAPITAL LETTER I
004A	J	LATIN CAPITAL LETTER J
004B	K	LATIN CAPITAL LETTER K
004C	L	LATIN CAPITAL LETTER L
004D	M	LATIN CAPITAL LETTER M
004E	N	LATIN CAPITAL LETTER N
004F	O	LATIN CAPITAL LETTER O
0050	P	LATIN CAPITAL LETTER P
0051	Q	LATIN CAPITAL LETTER Q
0052	R	LATIN CAPITAL LETTER R
0053	S	LATIN CAPITAL LETTER S
0054	T	LATIN CAPITAL LETTER T
0055	U	LATIN CAPITAL LETTER U
0056	V	LATIN CAPITAL LETTER V
0057	W	LATIN CAPITAL LETTER W
0058	X	LATIN CAPITAL LETTER X
0059	Y	LATIN CAPITAL LETTER Y
005A	Z	LATIN CAPITAL LETTER Z
005B	[	LEFT SQUARE BRACKET
005C	\	REVERSE SOLIDUS
005D	]	RIGHT SQUARE BRACKET
005E	^	CIRCUMFLEX ACCENT
005F	_	LOW LINE
0060	`	GRAVE ACCENT
0061	a	LATIN SMALL LETTER A

0062	b	LATIN SMALL LETTER B
0063	c	LATIN SMALL LETTER C
0064	d	LATIN SMALL LETTER D
0065	e	LATIN SMALL LETTER E
0066	f	LATIN SMALL LETTER F
0067	g	LATIN SMALL LETTER G
0068	h	LATIN SMALL LETTER H
0069	i	LATIN SMALL LETTER I
006A	j	LATIN SMALL LETTER J
006B	k	LATIN SMALL LETTER K
006C	l	LATIN SMALL LETTER L
006D	m	LATIN SMALL LETTER M
006E	n	LATIN SMALL LETTER N
006F	o	LATIN SMALL LETTER O
0070	p	LATIN SMALL LETTER P
0071	q	LATIN SMALL LETTER Q
0072	r	LATIN SMALL LETTER R
0073	s	LATIN SMALL LETTER S
0074	t	LATIN SMALL LETTER T
0075	u	LATIN SMALL LETTER U
0076	v	LATIN SMALL LETTER V
0077	w	LATIN SMALL LETTER W
0078	x	LATIN SMALL LETTER X
0079	y	LATIN SMALL LETTER Y
007A	z	LATIN SMALL LETTER Z
007B	{	LEFT CURLY BRACKET
007C		VERTICAL LINE
007D	}	RIGHT CURLY BRACKET
007E	~	TILDE

[Приложение 5 следует]

Приложение 5

Дополнительные требования для обмена данными (только для патентных ведомств)

Версия 1.1

*Одобрено Комитетом по Стандартам ВОИС (КСВ)
на своей пятой сессии 2 июня 2017*

В контексте обмена данными с базами данных провайдеров, патентные ведомства для каждой последовательности должны заполнить элемент INSD_other-seqids с одним INSDSeqid, содержащим ссылку на соответствующий опубликованный патент и идентификационный номер последовательности в следующем формате:

pat{office code}{publication number}{document kind code}{sequence identification number}

где {office code} - это код ведомства промышленной собственности, публикующего патентный документ в соответствии со стандартом ST.3; {document kind code} - это код для идентификации различных видов патентных документов, как указано в стандарте ST.16; {publication number} - это номер публикации заявки или патента; и {Sequence identification number} - это номер последовательности в данной заявке или патенте.

Пример:

pat|WO|2013999999|A1|123456.

Который будет преобразован в настоящий XML пример как:

```
<INSDSeq_other-seqids>  
  <INSDSeqid> pat|WO|2013999999|A1|123456</INSDSeqid>  
</INSDSeq_other-seqids>
```

Где “123456” это 123456-я последовательность из заявки WO номер 2013999999 (A1).

[Конец Приложения 5 и Стандарта]