

Comité de Normas Técnicas de la OMPI (CWS)

Quinta Sesión

Ginebra, 29 de mayo a 2 de junio de 2017

REVISIÓN DE LA NORMA ST.26 DE LA OMPI

Documento preparado por la Secretaría

1. En la reanudación su cuarta sesión, celebrada en Ginebra en marzo de 2016, el Comité de Normas Técnicas de la OMPI (CWS) adoptó la Norma ST.26 de la OMPI “Norma recomendada para la presentación de listas de secuencias de nucleótidos y aminoácidos en lenguaje extensible de marcado (XML)”. En consecuencia, el CWS dio por buena la modificación de la descripción de la Tarea No. 44: “Preparar recomendaciones para las disposiciones relativas a la transición de la Norma ST.25 a la Norma ST.26 de la OMPI; y una propuesta para la revisión de la Norma ST.26 de la OMPI, de ser necesario.” (Véanse los párrafos 49 a 53 y 82 a 84 del documento CWS/4BIS/16.)
2. En el marco de la nueva descripción de la Tarea No. 44, el Equipo Técnico SEQL ha llevado a cabo debates sobre la revisión de la Norma ST.26 de la OMPI; la Oficina Europea de Patentes (OEP), en calidad de responsable del Equipo Técnico, proporcionó un informe sobre la marcha de la labor realizada por el Equipo Técnico, que se reproduce en el Anexo I del presente documento.
3. Como resultado de sus deliberaciones, el Equipo Técnico SEQL sometió a consideración y, en caso apropiado, aprobación del CWS, la propuesta definitiva de revisión de la Norma ST.26, que contiene modificaciones del cuerpo principal de la norma y sus Anexos I a III y un nuevo Anexo VI (Documento de orientación). Cabe señalar que los Anexos IV y V de la Norma ST.26 permanecen sin cambios (por lo tanto, no se adjuntan). La propuesta de revisión se reproduce en el Anexo II del presente documento con respecto al cuerpo principal de la Norma ST.26 y sus Anexos I a II (titulados “Anexo I de la Norma ST.26” y “Anexo II de la Norma ST.26”, respectivamente); en cuanto al Anexo III de la Norma ST.26, el Equipo Técnico SEQL propone cambiar los códigos de dos letras para el idioma de mayúsculas a minúsculas a fin de que se

hallen en concordancia con los códigos de dos letras definidos en la Norma ISO 639, por ejemplo, cambiar el código de dos letras “EN” para el inglés por “en”. Cabe señalar que si el CWS aprueba la propuesta de revisión, la nueva versión de la Norma ST.26 sería la versión 1.1 (el Anexo III de la Norma ST.26 no se adjunta al presente documento).

4. *Se invita al CWS a:*

(a) *tomar nota del contenido del presente documento y del informe sobre la labor del Equipo Técnico SEQL que figura en el Anexo I del presente documento; y*

(b) *considerar la aprobación de la propuesta de revisión de la Norma ST.26 de la OMPI, mencionada en el párrafo 3 y reproducida en el Anexo II del presente documento, y tomar una decisión al respecto.*

[Sigue el Anexo I]

INFORME SOBRE LA REVISIÓN DE LA NORMA ST.26 DE LA OMPI POR EL EQUIPO TÉCNICO SEQL

Documento preparado por la Oficina Europea de Patentes (OEP)

ANTECEDENTES

1. El Equipo Técnico sobre listas de secuencias fue creado por el Comité de Normas Técnicas de la OMPI (CWS), en su primera sesión (25 a 29 de octubre de 2010), para que se ocupara de la Tarea No. 44 (véase el párrafo 29 del documento CWS/1/10):
2. “Preparar una recomendación sobre la presentación de listas de secuencias de nucleótidos y aminoácidos en lenguaje extensible de marcado (XML) para su adopción como norma técnica de la OMPI. La propuesta de una nueva norma técnica de la OMPI se presentará junto con un informe sobre la incidencia de dicha norma en la actual Norma ST.25 de la OMPI, incluyendo propuestas de los cambios que sea necesario introducir en la Norma ST.25”.
3. También se solicitó al Equipo Técnico que:
“[estableciera] contacto con el órgano pertinente del PCT teniendo en cuenta la posible incidencia de la nueva norma en el Anexo C de las Instrucciones Administrativas del PCT.”
4. Se encargó a la Oficina Europea de Patentes (OEP) la función de responsable del Equipo Técnico, y la nueva norma técnica de la OMPI, titulada Norma ST.26, fue adoptada formalmente durante la reanudación de la cuarta sesión del CWS (CWS/4BIS) en marzo de 2016.
5. En la reanudación de su cuarta sesión (21 a 24 de marzo de 2016), el CWS aprobó la descripción modificada de la Tarea No. 44, que reza actualmente:
“Preparar recomendaciones para las disposiciones relativas a la transición de la Norma ST.25 a la Norma ST.26 de la OMPI; y una propuesta para la revisión de la Norma ST.26 de la OMPI, de ser necesario.”

INFORME SOBRE LA MARCHA DE LA LABOR

6. Tras la adopción de la Norma ST.26 en la reanudación de la cuarta sesión del CWS, el Equipo Técnico llevó a cabo dos rondas de deliberaciones, la ronda 8, centrada en la manera de seguir mejorando la norma adoptada a fin de velar por su eficacia una vez implementada, y la ronda 9, sobre cuestiones relativas la transición (de la Norma ST.25 a la Norma ST.26). Durante ese período, el Equipo Técnico celebró varias reuniones por videoconferencia (WebEx) y dos reuniones oficiosas en Ginebra (23 de marzo y 9 de septiembre de 2016).
7. El Equipo Técnico se dedicó en concreto a la revisión de la Norma ST.26 adoptada, con el fin de facilitar su aplicación por los solicitantes y las Oficinas. La Norma ST.26 revisada, sometida a la quinta sesión del CWS para su aprobación, comprende los siguientes cambios principales:
 - aclaración de los ácidos nucleicos péptidos (PNA) y de las variantes de secuencias en la Norma;
 - incorporación de un documento de orientación (Anexo VI) para facilitar la armonización de las prácticas y la interpretación entre las Oficinas y los solicitantes;
 - actualización del Anexo I (Vocabulario controlado) a efectos de su concordancia con la versión 10.6 del cuadro de características del INSDC, que fue publicado en noviembre de 2016;

- añadido/nueva redacción de comentarios en el Anexo II (DTD) a efectos de aclaración y concordancia con la versión 1.5 de las definiciones de tipo de documento del INSDC.
- mejora del texto general de la Norma con arreglo a las consultas públicas celebradas por la Oficina Europea de Patentes, la Oficina Japonesa de Patentes y la Oficina de Patentes y Marcas de los Estados Unidos en 2016/2017.

HOJA DE RUTA

8. En la quinta sesión del CWS, obtener la aprobación de la revisión/añadido de los textos de la Norma ST.26.
9. En la quinta sesión del CWS, obtener la aprobación de las "Recomendaciones para las disposiciones relativas a la transición de la Norma ST.25 a la Norma ST.26 de la OMPI".
10. Prestar apoyo a la Oficina Internacional de la OMPI facilitando los requisitos de los usuarios y comentarios sobre la herramienta de autoría.
11. Prestar apoyo a la Oficina Internacional en la consiguiente revisión de las Instrucciones Administrativas del PCT.
12. Dedicarse a la revisión de la Norma ST.26 de la OMPI en el futuro. Se sugiere que, en adelante, las revisiones de la norma sean propiciadas por miembros del CWS, en lugar de obedecer a un calendario predeterminado.

[Sigue el Anexo II]

NORMA ST.26

NORMA RECOMENDADA PARA LA PRESENTACIÓN DE LISTAS DE SECUENCIAS DE NUCLEÓTIDOS Y AMINOÁCIDOS MEDIANTE EL LENGUAJE EXTENSIBLE DE MARCADO (XML)

Versión ~~1.0~~1.1

*Propuesta presentada por el Equipo Técnico SEQL para su consideración y aprobación en la quinta sesión del CWS
Adoptada por el Comité de Normas de la OMPI (CWS)
en la reanudación de la cuarta sesión el 24 de marzo de 2016*

Nota editorial de la Oficina Internacional

El Comité de Normas Técnicas de la OMPI (CWS) acuerda pedir a las oficinas de propiedad industrial que aplacen los preparativos para la aplicación de la nueva Norma ST.26 de la OMPI hasta que el CWS apruebe las recomendaciones para la transición de la Norma ST.25 a la nueva Norma ST.26 en su siguiente sesión, que se debería celebrar en 2017. Mientras tanto, se deberá seguir usando la Norma ST.25.

La Norma se publica a los fines de la información de las oficinas de propiedad industrial y otras partes interesadas.

ÍNDICE

INTRODUCCIÓN	3
DEFINICIONES.....	3
ÁMBITO	4
REFERENCIAS.....	5
PRESENTACIÓN REPRESENTACIÓN DE SECUENCIAS	5
<i>Secuencias de nucleótidos</i>	5
<i>Secuencias de aminoácidos</i>	8
<i>Presentación de casos especiales</i>	10
ESTRUCTURA DE LA LISTA DE SECUENCIAS EN FORMATO XML.....	10
<i>Elemento raíz</i>	11
<i>Parte de información general</i>	11
<i>Parte de datos de secuencia</i>	15
<i>Cuadro de características</i>	16
<i>Claves de caracterización</i>	17
<i>Claves de caracterización obligatorias</i>	17
<i>Localización de característica</i>	17
<i>Calificadores de caracterización</i>	19
<i>Calificadores de caracterización obligatorios</i>	19
<i>Elementos de los calificadores</i>	19
<i>Texto libre</i>	21
<i>Secuencias codificadoras</i>	21
<i>Variantes</i>	22

ANEXOS

Anexo I - Vocabulario controlado

Anexo II - Definiciones de tipo de documento (DTD) para las listas de secuencias

Anexo III - Ejemplo de lista de secuencias (archivo XML)

Anexo IV - Subconjunto de caracteres del cuadro de códigos de caracteres del alfabeto latino básico de la norma Unicode

Anexo V - Requisitos adicionales sobre el intercambio datos (únicamente para las oficinas de patentes)

Anexo VI – Documento de orientación

Apéndice

NORMA ST.26

NORMA RECOMENDADA PARA LA PRESENTACIÓN DE LISTAS DE SECUENCIAS DE NUCLEÓTIDOS Y AMINOÁCIDOS MEDIANTE EL LENGUAJE EXTENSIBLE DE MARCADO (XML)

Versión 1.01.1

*Propuesta presentada por el Equipo Técnico SEQL para su consideración y aprobación en la quinta sesión del CWS
Adoptada por el Comité de Normas de la OMPI (CWS)
en la reanudación de la cuarta sesión el 24 de marzo de 2016*

INTRODUCCIÓN

1. Esta Norma define la manera de divulgar en una solicitud de patente las secuencias de nucleótidos y aminoácidos que deben figurar en una lista de secuencias, ~~las características~~ **la representación** de esas divulgaciones, y la definición de tipo de documento (DTD) cuando las listas de secuencias se presentan en lenguaje extensible de marcado (XML). Se recomienda a las oficinas de propiedad industrial que acepten toda lista de secuencias compatible con esta Norma, que se presente en una solicitud de patente o en relación con una solicitud de patente.

2. La Norma tiene por objetivo:

- (a) permitir que el solicitante establezca una única lista de secuencias en una solicitud de patente que sea aceptable a los efectos tanto de los procedimientos internacionales como nacionales o regionales;
- (b) mejorar la precisión y calidad de la presentación de las secuencias a fin de facilitar su difusión para beneficio de los solicitantes, el público y los examinadores;
- (c) facilitar la búsqueda de datos en las secuencias; y
- (d) permitir el intercambio electrónico de datos sobre las secuencias y la introducción de esos datos en bases de datos informatizadas.

DEFINICIONES

3. A los efectos de la presente Norma:

(a) por "aminoácido" se entenderá todo aminoácido que pueda ser representado mediante cualquiera de los símbolos descritos en el Anexo I (véase el Cuadro 3 de la Sección 3). Quedan comprendidos entre tales aminoácidos, los D-aminoácidos y los aminoácidos que contienen cadenas laterales modificadas o sintéticas. Los aminoácidos deberán interpretarse como L-aminoácidos no modificados a menos que se indique con detalle **en el cuadro de características** que se trata de aminoácidos modificados tal como se prevé en el párrafo ~~2930~~. **A los fines de la presente norma, un residuo de ácido nucleico péptido (PNA) no se considera aminoácido, pero se considera nucleótido según se describe en el párrafo 3.q)i)2).**

(b) por "vocabulario controlado" se entenderá la terminología descrita en la presente Norma que deberá utilizarse a la hora de indicar las características de una secuencia, a saber, las anotaciones de regiones o sitios de interés tal como figuran en el Anexo I.

(c) por "enumeración de sus residuos" se entenderá la divulgación de una secuencia en una solicitud de patente en la que se enumera, por orden, cada residuo de la secuencia, en la que:

- (i) **el residuo se representa mediante un nombre, abreviatura, símbolo o estructura (por ejemplo, HHHHHHQ o HisHisHisHisHisHisGln); o**
- (ii) **varios residuos se representan mediante una fórmula abreviada (por ejemplo, His₆Gln).**

(d) por "secuencia ignorada deliberadamente", o secuencia vacía, se entenderá un espacio reservado para mantener la numeración de las secuencias que figuran en la lista de secuencias a fin de garantizar su coherencia con la numeración de la divulgación, por ejemplo, para no tener que volver a numerar las secuencias contenidas en la divulgación y en la lista de secuencias cuando se suprime una secuencia de la divulgación.

(e) por "aminoácido modificado" se entenderá todo aminoácido descrito en el párrafo 3.a) distinto de L-alanina, L-arginina, L-asparagina, L-ácido aspártico, L-cisteína, L-glutamina, L-ácido glutámico, L-glicina, L-histidina, L-isoleucina, L-leucina, L-lisina, L-metionina, L-fenilalanina, L-prolina, L-pirrolisina, L-serina, L-selenocisteína, L-threonine, L-triptófano, L-tirosina, o L-valina.

(f) por "nucleótido modificado" se entenderá todo nucleótido ~~o~~ descrito en el párrafo 3.g) distinto de 3'-monofosfato de desoxiadenosina, 3'-monofosfato de desoxiguanosina, 3'-monophosphate de desoxicitidina, 3'-monofosfato de desoxitimidina, 3'-monofosfato de adenosina, 3'-monofosfato de guanosina, 3'-monofosfato de citidina o 3'-monofosfato de uridina.

(g) por "~~análogo de~~ nucleótido" **se entenderá todo nucleótido o análogo de nucleótido** que pueda representarse utilizando cualquiera de los símbolos descritos en el Anexo I (véase el Cuadro 1 de la Sección 1). ~~Los nucleótidos podrán contener, entre otros elementos,~~ **en el que el nucleótido o análogo de nucleótido contiene:**

(i) un componente del esqueleto seleccionado de:

- 1) 5' monofosfato de 2' desoxirribosa (el componente del esqueleto de un desoxirribonucleótido) o 5' monofosfato de ribosa (el componente de la estructura de un ribonucleótido); o
- 2) el análogo de un 5' monofosfato de 2' desoxirribosa o 5' monofosfato de ribosa, que al formar el esqueleto de un análogo de ácido nucléico, da lugar a una disposición de las nucleobases que imita la disposición de las nucleobases en ácidos nucleicos que contienen un esqueleto de 5' monofosfato de 2' desoxirribosa o 5' monofosfato de ribosa, en la que el análogo de ácido nucléico es capaz de aparear las bases con un ácido nucléico complementario; entre los ejemplos de análogos de nucleótidos figuran los aminoácidos en los ácidos peptidonucleicos, las moléculas de ácido glicólico en los ácidos nucleicos glicólicos, las moléculas de azúcar de treofuranosil en los ácidos nucleicos treósicos, los anillos de morfolina y los grupos de fosfordiamidata en los morfolinos, y las moléculas de ciclohexenilo en los ácidos nucleicos que contienen ciclohexeno.

o

(ii) el componente del esqueleto o bien:

- 1) está adherido a una nucleobase, incluida una nucleobase modificada o sintética de purina o pirimidina una base modificada o sintética de purina o pirimidina, o una ribosa o desoxirribosa modificada o sintética, y podrán unirse mediante un enlace internucleósido modificado o sintético en el sentido de 3' a 5', es decir, toda fracción química que ofrezca la misma función estructural que la fracción fosfato de ADN o ARN, como en el caso de una fracción de fosforotioato; o bien
- 2) al carecer de una nucleobase de purina o pirimidina cuando el nucleótido es parte de una secuencia nucleótida, se menciona como "sitio AP" o "sitio abásico".

(h) por "residuo" se entenderá todo nucleótido o aminoácido individual o sus análogos respectivos en una secuencia.

(i) por "identificador de secuencia" se entenderá un número entero único que se asigna a cada secuencia de la lista de secuencias.

(j) por "lista de secuencias" se entenderá una parte de la descripción de la solicitud de patente tal como fue presentada, o un documento presentado posteriormente a la solicitud, que presenta incluye la(s) secuencia(s) divulgada(s) de nucleótidos y/o aminoácidos junto con toda otra descripción detallada, según estipula la presente Norma.

(k) por "específicamente definido" se entenderá todo nucleótido distinto a los representados por el símbolo "n" y todo aminoácido distinto a los representados por el símbolo "X" que se enumeran en el Anexo I (véanse la Sección 1, Cuadro 1, y la Sección 3, Cuadro 3, respectivamente).

(l) por nucleótido o aminoácido "desconocido" se entenderá la presencia de un único nucleótido o aminoácido cuya identidad es desconocida o no se divulga.

4. A los fines de la presente Norma, el término o términos:

- (a) "podrá" hace referencia a un enfoque facultativo o autorizado, pero no a un requisito.
- (b) "deberá" hace referencia a un requisito de la Norma; la no observancia del requisito, dará lugar al incumplimiento.
- (c) "no deberá" hace referencia a una prohibición de la Norma.
- (d) "debería" hace referencia a un enfoque al que se insta vivamente, pero no a un requisito.
- (e) "no debería" hace referencia a un enfoque que se desaconseja vivamente, pero no a una prohibición.

ÁMBITO

5. La presente Norma establece los requisitos de presentación de las listas de secuencias de nucleótidos y aminoácidos de las secuencias divulgadas en las solicitudes de patente.

6. Una lista de secuencias compatible con la presente Norma (en adelante lista de secuencias) contiene una parte de información general y una parte de datos de secuencia. La lista de secuencias deberá presentarse en un único archivo en formato XML utilizando la definición de tipo de documento (DTD) descrita en el Anexo II. El objetivo de la información bibliográfica contenida en la parte de información general sirve únicamente para vincular la lista de secuencias a la solicitud de patente en el marco de la cual se presenta dicha lista de secuencias. La parte de datos de secuencia está compuesta por uno o más elementos de datos de secuencia, cada uno de los cuales contiene información acerca de una secuencia. Los elementos de datos de secuencia incluyen varias claves de caracterización y los calificadores subsiguientes basados en las especificaciones de la *International Nucleotide Sequence Database Collaboration* (INSDC) y UniProt.

7. A los efectos de la presente Norma, una secuencia deberá figurar en una lista de secuencias cuando se divulga en cualquier parte de una solicitud mediante la enumeración de sus residuos y ~~se trata de~~ puede representarse como:

(a) una secuencia no ramificada o una ~~porción~~ región lineal de una secuencia ramificada que contiene diez o más nucleótidos específicamente definidos, en la que los nucleótidos adyacentes están unidos ~~en el sentido 3' a 5' (o 5' a 3')~~, o mediante:

(i) un enlace fosfodiéster en el sentido 3' a 5' (o 5' a 3'); o

(ii) cualquier enlace químico que dé lugar a una disposición de nucleobases adyacentes que imite la disposición de las nucleobases en ácidos nucleicos que ocurren naturalmente; o

(b) una secuencia no ramificada o una ~~porción~~ región lineal de una secuencia ramificada que contiene cuatro o más aminoácidos específicamente definidos, en la que los aminoácidos adyacentes están unidos por enlaces peptídicos.

8. Una lista de secuencias no deberá incluir, en calidad de secuencia a la que se ha asignado su propio número de identificación, ninguna secuencia que tenga menos de diez nucleótidos específicamente definidos, o menos de cuatro aminoácidos específicamente definidos.

REFERENCIAS

9. Las siguientes Normas y recursos son referencias pertinentes para la presente Norma:

International Nucleotide Sequence Database Collaboration (INSDC)

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

Norma Internacional ISO 639-1:2002

Codes for the representation of names of languages Part 1: Alpha-2 code;

UniProt Consortium

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

W3C XML 1.0

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

Norma técnica de la OMPI [ST.2](#)

Forma normalizada de designar las fechas según el calendario gregoriano;

Norma técnica de la OMPI [ST.3](#)

Códigos normalizados de dos letras, recomendados para la representación de Estados, otras entidades y organizaciones intergubernamentales;

Norma técnica de la OMPI [ST.16](#)

Código normalizado para la identificación de los diferentes tipos de documentos de patente;

Norma técnica de la OMPI [ST.25](#)

Norma para la presentación de listas de secuencias de nucleótidos y aminoácidos en solicitudes de patente.

PRESENTACIÓN/REPRESENTACIÓN DE SECUENCIAS

10. Se deberá asignar a cada secuencia ~~comprendida en el párrafo 7~~ un identificador de secuencia distinto, ~~incluidas las secuencias que sean idénticas a una región de una secuencia más larga~~. Los identificadores de secuencias deberían comenzar con el número 1, e irán aumentando de forma consecutiva por números enteros. Si a un identificador de secuencia no correspondiese una secuencia, a saber, una secuencia ignorada deliberadamente, se deberá utilizar el código "000" en lugar de la secuencia (véase el párrafo 58). El número total de secuencias deberá indicarse en la lista de secuencias y deberá ser igual al número total de identificadores de secuencias, con independencia de si van seguidos de una secuencia o del código "000".

Secuencias de nucleótidos

11. Toda secuencia de nucleótidos sólo deberá representarse mediante una cadena única, en el sentido 5' a 3' y de izquierda a derecha, ~~o en el sentido de izquierda a derecha que imite el sentido 5' a 3'~~. Los valores 5' y 3' ~~o cualquier otro valor similar~~ no deberán ~~aparecer~~ estar incluidos en la secuencia. Toda secuencia de nucleótidos de doble cadena divulgada mediante la enumeración de los residuos de ambas cadenas deberá ~~presentarse~~ representarse de la siguiente manera:

(a) una única secuencia o dos secuencias distintas, a las que se les asignará su propio identificador de secuencia, en el que las dos cadenas distintas deberán ser plenamente complementarias entre sí, o

(b) dos secuencias distintas, a las que se les asignará su propio identificador de secuencia, en el que las dos cadenas no son plenamente complementarias entre sí.

12. ~~A los fines de la presente Norma, el primer nucleótido presentado en la secuencia será el residuo de la posición número 1. La numeración de las posiciones de nucleótidos deberá comenzar en la primera base de la secuencia con el número 1. Esa numeración deberá ser continua a lo largo de toda la secuencia en el sentido 5' a 3'.~~ Cuando las secuencias de nucleótidos tengan una configuración circular, el solicitante deberá ~~El método de numeración descrito para las secuencias de nucleótidos también es aplicable a las secuencias de nucleótidos de configuración circular. En ese caso, el solicitante deberá elegir el nucleótido en el que comienza la numeración.~~ del residuo de la posición número 1. La numeración será continua a lo largo de toda la secuencia en el sentido 5' a 3', o en el sentido que imite el sentido 5' a 3'. El último número de posición de los residuos deberá ser igual al número de nucleótidos de la secuencia.

13. Todos los nucleótidos de una secuencia deberán representarse mediante los símbolos descritos en el Anexo I (véase el Cuadro 1 de la Sección 1). Se deberá utilizar únicamente letras minúsculas. Todo símbolo utilizado para representar un nucleótido equivale a un único residuo.

14. El símbolo “t” se interpretará como timina en ADN y uracilo en ARN. El uracilo en ADN o la timina en ARN se considerará como un nucleótido modificado y deberá ir acompañado de una descripción detallada describirse detalladamente en el cuadro de características tal como se prevé en el párrafo 1819.

15. Si fuera necesario utilizar un símbolo de ambigüedad (que represente dos o más bases alternativas nucleótidos alternativos), debería utilizarse el símbolo más restrictivo, según figura en el Anexo I (sección 1, Cuadro 1). Por ejemplo, si una base un nucleótido en una determinada posición pudiera ser “a” o “g”, se debería utilizar “r”, en vez de “n”. El símbolo “n” se interpretará como “a”, “c”, “g”, o “t/u”, excepto cuando se utilice en una descripción detallada tal como se prevé en los párrafos 16 y 17 o 2021. El símbolo “n” no podrá deberá utilizarse para representar un elemento distinto a un nucleótido. El símbolo “n” podrá representar un único nucleótido modificado o “desconocido”, junto con una descripción detallada en el cuadro de características, tal como se prevé en los párrafos 16 y 17 o 2021. Para la representación de variantes de secuencias, por ejemplo, alternativas, supresiones, inserciones o sustituciones, véanse los párrafos 92 a 98.

16. Los nucleótidos modificados deberían representarse en la secuencia como los correspondientes bases nucleótidos no modificados, a saber, “a”, “c”, “g” o “t” cuando sea posible. Todo nucleótido modificado en una secuencia que no pueda representarse de otra manera por ningún otro símbolo descrito en el Anexo I (véase el Cuadro 1 de la Sección 1), por ejemplo, “otro” nucleótido, tal como un nucleótido que no existe en estado natural, deberá representarse mediante el símbolo “n”. Cuando se utiliza el símbolo “n” para representar un nucleótido modificado, éste equivale a un único residuo.

17. Los nucleótidos modificados deberán describirse con detalle en el cuadro de características (véanse los párrafos 5960 y siguientes) utilizando la clave de caracterización “modified_base” y el calificador obligatorio “mod_base”. El valor calificador deberá corresponder a una abreviatura única que figure en el Anexo I (véase el Cuadro 2 de la Sección 2; si la abreviatura es “OTHER”, el nombre completo no abreviado del nucleótido base modificado deberá indicarse como valor en un calificador “note”. Para una lista de nucleótidos modificados alternativos, podrá utilizarse el valor calificador “OTHER” junto con otro calificador “note” (véanse los párrafos 95 y 96). Las abreviaturas (o nombres completos) que figuran en el Anexo I (véase el Cuadro 2 de la Sección 2) a las que se ha referencia supra no deberán utilizarse en la propia secuencia.

18. Una secuencia de nucleótido que contenga uno o más regiones de nucleótidos modificados consecutivos que compartan el mismo componente del esqueleto (véase el párrafo 3.g)i)2)), deberá describirse detalladamente en el cuadro de características según se dispone en el párrafo 17. Los nucleótidos modificados de cada región podrán describirse conjuntamente en un único elemento INSDFeature según lo previsto en el párrafo 22. Deberá proporcionarse el nombre químico no abreviado más restrictivo que comprenda todos los nucleótidos modificados en la serie o una lista de los nombres químicos de todos los nucleótidos, como el valor del calificador “note”. Por ejemplo, una secuencia de ácido nucleico glicólico que contenga las nucleobases “a”, “c”, “g” o “t” podrá describirse en el calificador “note” como “2,3-dihydroxypropyl nucleosides.” Por otra parte, la misma secuencia podrá describirse en el calificador “note” como “2,3-dihydroxypropyladenine, 2,3-dihydroxypropylthymine, 2,3-dihydroxypropylguanine, or 2,3-dihydroxypropylcytosine.” Cuando un nucleótido modificado en la región incluya una modificación adicional, el nucleótido modificado deberá describirse detalladamente en el cuadro de características según lo previsto en el párrafo 17.

19. El uracilo en ADN o la timina en ARN se consideran nucleótidos modificados y deberán representarse en la secuencia por una “t” y describirse con detalle en el cuadro de características utilizando la clave de caracterización “modified_base”, el calificador “mod_base” con el valor calificador “OTHER” y el calificador “note” con el valor calificador “uracil” o “thymine”, respectivamente.

20. Los siguientes ejemplos ilustran la presentación representación de los nucleótidos modificados en la forma prevista en los párrafos 16 y 17 a 18:

Ejemplo 1: Nucleótido modificado utilizando una abreviatura que figura en el Anexo I (véase el Cuadro 2 de la Sección 2).

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

Ejemplo 2: Nucleótido modificado "xanthine" utilizando "OTHER" como figura en el Anexo I (véase el Cuadro 2 de la Sección 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>
```

Ejemplo 3: Una secuencia de nucleótido compuesta de nucleótidos modificados comprendidos en el párrafo 3.g)i)2) con dos nucleótidos individuales que incluyen otra modificación

```
<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. Todo nucleótido "desconocido" deberá representarse mediante el símbolo "n" en la secuencia. Los nucleótidos "desconocidos" deberían describirse con detalle en el cuadro de características (véanse los párrafos 60 siguientes) utilizando la clave de caracterización "unsure". El símbolo "n" equivale a un único residuo.

22. Toda región que contiene un número conocido de residuos contiguos "a", "c", "g", "t", o "n" para los cuales se aplica la misma descripción podrá describirse en conjunto utilizando un único elemento INSDFeature con la sintaxis "x.y" como descriptor de localización en el elemento INSDFeature_location (véanse los párrafos 6564 a 7271). Para la presentación de las variantes de secuencia, es decir, supresiones, inserciones o sustituciones, véanse los párrafos 92 a 9798.

23. El siguiente ejemplo ilustra la **representación** de una región de nucleótidos modificados para los cuales se aplica la misma descripción en la forma prevista en el párrafo 24.22:

```
<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>
```

Secuencias de aminoácidos

24. Los aminoácidos de una secuencia **proteínica o peptídica de aminoácidos** deberán **relacionarse** **representarse** en el sentido del grupo amino al grupo carboxilo, y de izquierda a derecha. Los grupos amino y carboxilo no deberán representarse en la secuencia.

25. **A los fines de la presente Norma, la numeración de las posiciones de aminoácidos deberá comenzar en el primer aminoácido de la secuencia, con será el residuo de la posición número 1, incluidos los aminoácidos que preceden a la proteína madura, por ejemplo, las presecuencias, las prosecuencias, y las preprosecuencias, así como las secuencias señal. Esa numeración deberá ser continua. Cuando las secuencias de aminoácidos tengan una configuración circular, el solicitante deberá elegir el aminoácido del residuo de la posición número 1. La numeración será continua** a lo largo de toda la secuencia en el sentido de amino a carboxilo.

26. Todos los aminoácidos de una secuencia deberán representarse mediante los símbolos descritos en el Anexo I (véase el Cuadro 3 de la Sección 3). Se deberán utilizar únicamente letras mayúsculas. Todo símbolo utilizado para representar un aminoácido equivale a un único residuo.

27. Si fuese necesario utilizar un símbolo de ambigüedad (que represente dos o más aminoácidos posibles), debería utilizarse el símbolo más restrictivo. Por ejemplo, si un aminoácido en una posición dada podría ser un ácido aspártico o asparagina, debería utilizarse el símbolo "B" en vez del símbolo "X". El símbolo "X" se interpretará como uno de los símbolos "A", "R", "N", "D", "C", "Q", "E", "G", "H", "I", "L", "K", "M", "F", "P", "O", "S", "U", "T", "W", "Y", o "V", excepto cuando se utilice en una descripción detallada **en el cuadro de características** tal como se prevé en los párrafos 28.29 a 30.31 o 31.32 a 33. El símbolo "X" no **podrá** **deberá** utilizarse para representar un elemento distinto a un aminoácido. Un único aminoácido podrá representarse mediante el símbolo "X", junto con una descripción detallada **en el cuadro de características tal como se prevé en los párrafos 28.29 a 30.31 o 31.32 a 33. Para la presentación** **representación** de las variantes de secuencia, es decir, **alternativas**, supresiones, inserciones, o sustituciones, véanse los párrafos 92 a 97.98.

28. Las secuencias de aminoácidos **divulgadas** separadas por **uno o más espacios en blanco** o símbolos internos de terminación, **representados** por ejemplo, **por** "Ter" o asterisco "*" o punto ".", **en una divulgación, o un espacio en blanco,** deberán **presentarse** **incluirse** como secuencias distintas en el caso de cada secuencia de aminoácidos que contenga al menos cuatro aminoácidos específicamente definidos y esté contemplada en el párrafo 67. **Cada** **A cada** secuencia distinta de este tipo deberá **presentarse en la lista de secuencias con asignarse** su propio identificador de secuencia, **utilizando únicamente los símbolos descritos en el Anexo I (véase el Cuadro 3 de la Sección 3). No deberán utilizarse** **incluirse** símbolos de terminación ni espacios en las secuencias presentadas en una lista de secuencias **(véase el párrafo 57).**

29. Los aminoácidos modificados, incluidos los D-aminoácidos, deberían representarse en la secuencia como los correspondientes aminoácidos no modificados, cuando sea posible. Todo aminoácido modificado que aparece en una secuencia y que no pueda ser representado por ningún otro símbolo que figure en el Anexo I (véase el Cuadro 3 de la Sección 3), **por ejemplo, "otro" aminoácido,** deberá representarse por el símbolo "X". El símbolo "X" equivale a un único residuo.

30. Los aminoácidos modificados deberán describirse con detalle en el cuadro de características (véanse los párrafos 60 y siguientes). **Cuando proceda, deberían utilizarse las claves de caracterización "CARBOHYD" o "LIPID" junto con el calificador "NOTE".** La clave de caracterización "MOD_RES" debería utilizarse para **las otras** modificaciones posttraduccionales de los aminoácidos junto con el calificador "NOTE"; de lo contrario, debería utilizarse la clave de caracterización "SITE" junto con el calificador "NOTE". El valor del calificador "NOTE" deberá ser ya sea una abreviatura descrita en el Anexo I (véase el Cuadro 4 de la Sección 4), o el nombre completo no abreviado del aminoácido modificado. Las abreviaturas descritas en el Cuadro 4 que se mencionan *supra* o los nombres completos no abreviados no deberán utilizarse en la propia secuencia.

31. Los siguientes ejemplos ilustran la **representación** de aminoácidos modificados en la forma prevista en el párrafo 29:

Ejemplo 1: Aminoácido modificado (modificación postraduccional).

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

Ejemplo 2: Aminoácido modificado (modificación no postraduccional).

```
<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>
```

Ejemplo 3: D-aminoácido.

```
<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. Todo aminoácido “desconocido” u “otro” no contemplado en el párrafo 28, deberá representarse mediante el símbolo “X” en la secuencia. El símbolo “X” equivale a un único residuo. Todo aminoácido “desconocido” designado mediante “X” deberá describirse con detalle en el cuadro de características (véanse los párrafos 60 y siguientes.) mediante la clave de caracterización “UNSURE” y facultativamente el calificador “NOTE”. Todo aminoácido “otro” designado como “X” deberá describirse con detalle mediante la clave de caracterización “SITE” o “MOD_RES”, según el caso, y el calificador “NOTE” con el nombre completo no abreviado del aminoácido “otro”. El símbolo “X” es el equivalente de un único residuo.

33. Los siguientes ejemplos ilustran la **representación** de aminoácidos “desconocidos” u “otros” en la forma prevista en los párrafos 31 y 32 supra el párrafo 32:

Ejemplo 1: Aminoácido “desconocido”.

```
<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>
```

Ejemplo 2: Aminoácido "otro".

```
<INSDFeature>  
  <INSDFeature_key>SITE</INSDFeature_key>  
  <INSDFeature_location>3</INSDFeature_location>  
  <INSDFeature_qualis>  
    <INSDQualifier>  
      <INSDQualifier_name>NOTE</INSDQualifier_name>  
      <INSDQualifier_value>Homoserine</INSDQualifier_value>  
    </INSDQualifier>  
  </INSDFeature_qualis>  
</INSDFeature>
```

34. Toda región que contiene un número desconocido de residuos "X" contiguos para los cuales se aplica la misma descripción podrá describirse en conjunto utilizando la sintaxis "x.." como descriptor de localización en el elemento INSDFeature_location (véanse los párrafos 6564 a 7470). Para la **presentación** de variantes de secuencia, a saber, supresiones, inserciones, o sustituciones, véanse los párrafos 9293 a 9798.

Presentación de casos especiales

35. Toda secuencia divulgada mediante la enumeración de sus residuos, que se interpreta como una única secuencia contigua a partir de uno o más segmentos no contiguos de una secuencia más grande o de segmentos de diferentes secuencias, deberá figurar en la lista de secuencias **como una única secuencia con un único y se le deberá asignar su propio** identificador de secuencia.

36. Toda secuencia divulgada mediante la enumeración de sus residuos que contenga regiones de residuos específicamente **enumerados definidos** separados por una o más regiones de residuos "n" o "X" contiguos (véanse los párrafos 15 y 2627, respectivamente), en la que el número exacto de residuos "n" o "X" de cada región se divulga, deberá figurar en la lista de secuencias de una única secuencia **y se le deberá asignar su propio** con un único identificador de secuencia.

37. Toda secuencia divulgada mediante la enumeración de sus residuos que contenga regiones de residuos específicamente **enumerados definidos** y separados por uno o más huecos compuestos por un número desconocido o no divulgado de residuos **no deberá figurar en la lista de secuencias como una única secuencia. Cada región de residuos específicamente definidos que está comprendida en el párrafo 7 deberá** figurar en la lista de secuencias como varias secuencias distintas. Cada una de estas secuencias distintas deberá contener una región de residuos **específicamente enumerados con y se le deberá asignar** su propio identificador de secuencia, en el cual el número de secuencias distintas es igual al número de regiones de residuos específicamente enumerados. Las secuencias que contienen huecos de un número desconocido o no divulgado de residuos **no deberán figurar en la lista de secuencias como una única secuencia.**

ESTRUCTURA DE LA LISTA DE SECUENCIAS EN FORMATO XML

38. Según lo previsto en el párrafo 56 *supra*, la instancia XML de un archivo que contiene una lista de secuencias compatible con la presente Norma se compone de:

- (a) una parte de información general, que contiene la información relativa a la solicitud de patente a la que está asociada la lista de secuencias; y
- (b) una parte de datos de secuencia, que contiene uno o más elementos de datos de secuencia, cada uno de los cuales, a su vez contiene información acerca de una secuencia.

En el Anexo III se presenta un ejemplo de una lista de secuencias.

39. La lista de secuencias deberá presentarse en formato XML 1.0 utilizando la DTD presentada en el Anexo II "Definición de tipo de documento para listas de secuencias".

- (a) La primera línea de la instancia XML deberá contener la declaración XML siguiente:

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

- (b) La segunda línea de la instancia XML deberá contener una declaración de tipo de documento (DOCTYPE):

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

40. La lista de secuencias electrónica completa deberá figurar en un solo archivo. El archivo deberá cifrarse utilizando el lenguaje Unicode UTF-8, con las siguientes restricciones:

- (a) la información contenida en los elementos ApplicantName, InventorName e InventionTitle de la parte de información general podrá estar compuesta por cualquier carácter Unicode, excepto los caracteres reservados que deberán sustituirse como se describe en el párrafo 41; y
- (b) la información contenida en todos los demás elementos de la parte de información general y en todos los elementos de la parte de datos de secuencia
 - o deberá estar compuesta por caracteres imprimibles (incluido el carácter de espacio) del cuadro de códigos de caracteres del alfabeto latino básico de la norma Unicode excepto los caracteres reservados, que deberán sustituirse como se describe en el párrafo 41, (es decir, limitados a los puntos de código Unicode 0020, 0021, 0023 a 0026, 0028 a 003B, 003D, y 003F a 007E – véase el Anexo IV), y
 - o las únicas entidades de caracteres permitidas son las entidades predefinidas descritas en el párrafo 41.

41. En una instancia XML de una lista de secuencias, los siguientes caracteres reservados deberán sustituirse por las entidades predefinidas correspondientes cuando se utilicen en el valor de un atributo o el contenido de un elemento:

Carácter reservado	Entidades predefinidas
<	<
>	>
&	&
"	"
'	'

Véase el párrafo 7271 para consultar un ejemplo.

42. Todos los elementos obligatorios deberán indicarse (excepto en el caso definido en el párrafo 58 para las secuencias ignoradas deliberadamente). Los elementos facultativos para los cuales no existe ningún contenido disponible no deberían aparecer en la instancia XML (salvo lo dispuesto en el párrafo 95 para la representación de una supresión en una secuencia en el valor correspondiente al calificador "replace").

Elemento raíz

43. El elemento raíz de una instancia XML según lo dispuesto en esta Norma es el elemento ST26SequenceListing, que tiene los siguientes atributos:

Atributo	Descripción	Obligatorio/Facultativo
dtdVersion	Versión de la DTD utilizada para crear este archivo en el formato "V#_#", p. ej. "V1_0".	Obligatorio
fileName	Nombre del archivo que contiene la lista de secuencias.	Facultativo
softwareName	Nombre del programa informático que generó este archivo.	Facultativo
softwareVersion	Versión del programa informático que generó este archivo.	Facultativo
productionDate	Fecha de producción del archivo que contiene la lista de secuencias (programa "AACC-MM-DD").	Facultativo

44. El siguiente ejemplo ilustra el elemento raíz ST26SequenceListing, y sus atributos, de una instancia XML en la forma prevista en el párrafo 43 supra:

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

*{...} represents the general information part and the sequence data part that have not been included in this example.

Parte de información general

45. Los elementos de la parte de información general se relacionan a la información relativa a la solicitud de patente, de la siguiente manera:

Elemento	Descripción	Obligatorio/ Facultativo
<p>ApplicationIdentification</p> <p>La ApplicationIdentification se compone de:</p> <p>IPOfficeCode</p> <p>ApplicationNumberText</p> <p>FilingDate</p>	<p>La identificación de la solicitud para la cual se presenta la lista de secuencias</p> <p>El código ST.3 de la oficina de presentación</p> <p>La identificación de la solicitud suministrada por la oficina de presentación (por ejemplo, PCT/IB2013/099999)</p> <p>La fecha de presentación de la solicitud de patente para la cual se presenta la lista de secuencias (formato ST.2 a "AACC-MM-DD", que utiliza 4 dígitos para representar el año civil, 2 dígitos el mes civil y 2 dígitos el número ordinal de un día dentro del mes civil, por ejemplo, 2015-01-31)</p>	<p>Obligatorio cuando una lista de secuencias se suministra en cualquier momento posterior a la asignación del número de solicitud</p> <p>Obligatorio</p> <p>Obligatorio</p> <p>Obligatorio cuando una lista de secuencias se suministra en cualquier momento posterior a la asignación de una fecha de presentación</p>
<p>ApplicantFileReference</p>	<p>Un identificador único asignado por el solicitante para identificar una solicitud específica, escrito en los caracteres descritos en el párrafo 40.b)</p>	<p>Obligatorio cuando una lista de secuencias se suministra en cualquier momento anterior a la asignación del número de solicitud; de lo contrario, Facultativo</p>
<p>EarliestPriorityApplicationIdentification</p>	<p>La identificación de la reivindicación de prioridad más antigua (también contiene IPOfficeCode, ApplicationNumberText y FilingDate, véase ApplicationIdentification <i>supra</i>)</p>	<p>Obligatorio cuando se reivindica la prioridad</p>
<p>ApplicantName</p>	<p>El nombre del primer solicitante mencionado escrito en los caracteres descritos en el párrafo 40.a). Este elemento contiene el atributo obligatorio languageCode descrito en el párrafo 47.</p>	<p>Obligatorio</p>
<p>ApplicantNameLatin</p>	<p>Si se escribe ApplicantName en caracteres distintos a los descritos en el párrafo 40.b), también deberá escribirse la traducción o transliteración del nombre del primer solicitante mencionado en los caracteres descritos en el párrafo 40.b)</p>	<p>Obligatorio cuando ApplicantName contiene caracteres no latinos</p>
<p>InventorName</p>	<p>Nombre del primer inventor mencionado escrito en los caracteres descritos en el párrafo 40.a). Este elemento contiene el atributo obligatorio languageCode descrito en el párrafo 47.</p>	<p>Facultativo</p>

Elemento	Descripción	Obligatorio/ Facultativo
InventorNameLatin	Si InventorName se escribe en caracteres distintos a los descritos en el párrafo 40.b), podrá también escribirse la traducción o transliteración del inventor mencionado en primer lugar en los caracteres descritos en el párrafo 40.b)	Facultativo
InventionTitle	Título de la invención escrita en los caracteres descritos en el párrafo 40.a) en el idioma de presentación. La traducción del título de la invención en otros idiomas podrá escribirse en los caracteres descritos en el párrafo 40.a) utilizando varios elementos InventionTitle. Este elemento contiene el atributo obligatorio languageCode descrito en el párrafo 48. El título de la invención deberá contener de preferencia dos a siete palabras.	Obligatorio en el idioma de presentación. Facultativo en los otros idiomas.
SequenceTotalQuantity	El número total de todas las secuencias que figuran en la lista de secuencias, incluidas las secuencias ignoradas deliberadamente (o secuencias vacías) (véase el párrafo 9).	Obligatorio

46. Los siguientes ejemplos ilustran la presentación de la parte de información general de la lista de secuencias en la forma prevista en el párrafo 45 *supra*:

Ejemplo 1: Lista de secuencias presentada antes de la asignación de la identificación de la solicitud y la fecha de presentación.

```
<?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>
```

*{...} representa la información pertinente para cada secuencia que no se ha incluido en este ejemplo.

Ejemplo 2: Lista de secuencias presentada después de la asignación de la identificación de la solicitud y fecha de presentación.

```
<?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_SEQ1.xml"
softwareName="SQL-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="ENen">GENOS Co., Inc.</ApplicantName>
  <InventorName languageCode="ENen">Keiko Nakamura</InventorName>
  <InventionTitle languageCode="ENen">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>
```

*{...} representa la información pertinente para cada secuencia que no se ha incluido en este ejemplo.

47. El nombre del solicitante y, facultativamente, el nombre del inventor deberán indicarse en el elemento ApplicantName o InventorName, respectivamente, ya que en general se mencionan en el idioma de presentación de la solicitud. El código de idioma adecuado (véase [la referencia en el párrafo 89 a la ISO 639-1:2002](#)) deberá indicarse en el atributo languageCode de cada elemento. Cuando el nombre del solicitante indicado contenga caracteres distintos a los del alfabeto latino, como se describe en el párrafo 40.b), también deberá indicarse una transliteración o traducción del nombre del solicitante en caracteres del alfabeto latino en el elemento ApplicantNameLatin. Cuando el nombre del inventor indicado contenga caracteres distintos a los del alfabeto latino, también podrá indicarse una transliteración o traducción del nombre del inventor en caracteres del alfabeto latino en el elemento InventorNameLatin.

48. El título de la invención deberá indicarse en el elemento InventionTitle en el idioma de presentación y también podrá indicarse en otros idiomas utilizando varios elementos InventionTitle (véase el cuadro del párrafo 45). El código de idioma adecuado (véase [la referencia en el párrafo 89 a la ISO 639-1:2002](#)) deberá indicarse en el atributo languageCode del elemento.

49. El siguiente ejemplo ilustra la presentación de los nombres y el título de la invención en la forma prevista en los párrafos 47 y 48 *supra*:

Ejemplo: El nombre del solicitante y el nombre del inventor se presentan en caracteres japoneses y latinos, y el título de la invención se presenta en japonés, inglés y francés.

```
<ApplicantName languageCode="JAja">出願製薬株式会社</ApplicantName>
<ApplicantNameLatin>Shutsugan Pharmaceuticals Kabushiki Kaisha</ApplicantNameLatin>
<InventorName languageCode="JAja">特許 太田</InventorName>
<InventorNameLatin>Taro Tokkyo</InventorNameLatin>
<InventionTitle languageCode="JAja">efg タンパク質のためのマウス abcd-1 遺伝子
</InventionTitle>
<InventionTitle languageCode="ENen">en"n">Mus musculus abcd-1 gene for efg protein
</InventionTitle>
<InventionTitle languageCode="FRfr">fr">Gène abcd-1 de Mus musculus pour protéine
efg </InventionTitle>
```

Parte de datos de secuencia

50. La parte de datos de secuencia deberá constar de uno o más elementos `SequenceData`, y cada elemento contendrá información sobre una sola secuencia.

51. Cada elemento `SequenceData` deberá tener un atributo obligatorio `sequenceIDNumber`, el cual contiene el identificador de secuencia (véase el párrafo 910) de cada secuencia. Por ejemplo:

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

52. El elemento `SequenceData` deberá contener un elemento dependiente `INSDSeq`, que consiste en otros elementos dependientes de la siguiente manera:

Elemento	Descripción	Obligatorio/No incluido	
		Secuencias	Secuencias ignoradas deliberadamente
<code>INSDSeq_length</code>	Longitud de la secuencia	Obligatorio	Obligatorio, sin ningún valor
<code>INSDSeq_moltype</code>	Tipo de molécula	Obligatorio	Obligatorio, sin ningún valor
<code>INSDSeq_division</code>	Indicación de que una secuencia está asociada a una solicitud de patente	Obligatorio con el valor "PAT"	Obligatorio, sin ningún valor
<code>INSDSeq_feature-table</code>	Lista de anotaciones de la secuencia	Obligatorio	NO debe incluirse
<code>INSDSeq_sequence</code>	Secuencia	Obligatorio	Obligatorio con el valor "000"

53. El elemento `INSDSeq_length` deberá divulgar el número de nucleótidos o aminoácidos de la secuencia contenidos en el elemento `INSDSeq_sequence`. Por ejemplo:

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

54. El elemento `INSDSeq_moltype` deberá divulgar el tipo de molécula que se está **presentando representando**. En el caso de las secuencias de nucleótidos, **incluidas las secuencias de análogos de nucleótidos**, el tipo de molécula deberá indicarse como ADN o ARN. Para las secuencias **proteínicas o polipeptídicas de aminoácidos**, el tipo de molécula deberá indicarse como AA. (Este elemento es distinto de los calificadores "mol_type" y "MOL_TYPE" descritos en los párrafos 55 y 8584). Por ejemplo:

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

55. Si una secuencia de nucleótidos contiene fragmentos ADN y ARN, el valor de `INSDSeq_moltype` deberá ser "DNA". La molécula combinada ADN/ARN deberá describirse con detalle en el cuadro de características, utilizando la clave de caracterización "source" y el calificador obligatorio "organism" con el valor "synthetic construct" y el calificador obligatorio "mol_type" con el valor "other DNA". Cada fragmento de ADN y ARN de la molécula combinada ADN/ARN debería describirse con detalle mediante la clave de caracterización "misc_feature" y el calificador "note", lo que indica si el fragmento es un ADN o un ARN.

56. El siguiente ejemplo ilustra la descripción de una secuencia de nucleótidos que contiene fragmentos tanto ADN como ARN, en la forma prevista en el párrafo 55 *supra*:

```
<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>
  </INSDSeq>
```

```

<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>
  cgaccacgcgtccgaggaaccaaccatcacgtttgaggacttcgtgaaggaattggataatacccgtcctaccaaagt
  gcgagcgccgactcattgctcctcgtaccgctcgagcggc
</INSDSeq_sequence>
</INSDSeq>

```

57. El elemento INSDSeq_sequence deberá divulgar la secuencia. Los residuos de la secuencia deberán presentarse de forma contigua utilizando únicamente Solo deberán incluirse en la secuencia los símbolos adecuados descritos en el Anexo I (véase el Cuadro 1 de la Sección 1 y el Cuadro 3 de la Sección 3). La secuencia no deberá contener incluir números, signos de puntuación o caracteres en blanco.

58. Toda secuencia ignorada deliberadamente deberá presentarse incluirse en la lista de secuencias y representarse de la siguiente manera:

- (a) el elemento SequenceData y su atributo sequenceIDNumber, indicando como valor el identificador de secuencia de la secuencia ignorada;
- (b) los elementos INSDSeq_length, INSDSeq_moltype, INSDSeq_division, presentes pero sin indicar ningún valor;
- (c) el elemento INSDSeq_feature-table no deberá incluirse; y
- (d) el elemento INSDSeq_sequence, con el valor "000".

59. El siguiente ejemplo ilustra la presentación representación de una secuencia ignorada deliberadamente en la forma prevista en el párrafo 58 *supra*:

```

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

```

Cuadro de características

60. El cuadro de características contiene información sobre la localización y las funciones de varias regiones de una secuencia específica. Se requiere un cuadro de características para cada secuencia, excepto para toda secuencia ignorada deliberadamente, en cuyo caso no deberá figurar. El cuadro de características figura en el elemento INSDSeq_feature-table, que consiste en uno o más elementos INSDFeature.

61. Cada elemento `INSDFeature` describe una característica, y consiste en elementos dependientes de la siguiente manera:

Elemento	Descripción	Obligatorio/Facultativo
<code>INSDFeature_key</code>	Palabra o abreviatura que indica una característica	Obligatorio
<code>INSDFeature_location</code>	Región de la secuencia presentada que corresponde a la característica	Obligatorio
<code>INSDFeature_qual</code>	Calificador que contiene información auxiliar acerca de la característica	Obligatorio si la clave de caracterización requiere uno o más calificadores, por ejemplo, <code>source</code> ; de lo contrario, Facultativo

Claves de caracterización

62. El Anexo I contiene la lista completa de claves de caracterización que deberán utilizarse en virtud de esta Norma, junto con la lista completa de los calificadores asociados y una indicación sobre el carácter obligatorio o facultativo de esos calificadores. La Sección 5 del Anexo I presenta la lista completa de las claves de caracterización para las secuencias de nucleótidos y la Sección 7 presenta la lista completa de las claves de caracterización para las secuencias de aminoácidos.

Claves de caracterización obligatorias

63. La clave de caracterización `"source"` es obligatoria para todas las secuencias de nucleótidos y la clave de caracterización `"SOURCE"` es obligatoria para todas las secuencias de aminoácidos, excepto para toda secuencia ignorada deliberadamente. Cada secuencia deberá tener una única clave de caracterización `"source"` o `"SOURCE"` que abarque toda la secuencia. Si una secuencia proviene de varias fuentes, esas fuentes podrán describirse con detalle en el cuadro de características, utilizando la clave de caracterización `"misc_feature"` y el calificador `"note"` para las secuencias de nucleótidos, y la clave de caracterización `"REGION"` y el calificador `"NOTE"` para las secuencias de aminoácidos.

64. ~~Algunas claves de caracterización requieren otra clave de caracterización, denominada "Parent Key", que debe utilizarse junto con esas claves de caracterización; por ejemplo, la clave de caracterización "C_region" requiere la clave de caracterización "CDS" (véase la Sección 5 del Anexo I).~~

Localización de característica

65. El elemento obligatorio `INSDFeature_location` deberá contener al menos un descriptor de localización, que defina un sitio o una región correspondiente a la característica de la secuencia en el elemento `INSDSeq_sequence`, y podrá contener uno o más operador(es) de localización (véanse los párrafos ~~6867~~ a ~~7470~~).

66. El descriptor de localización puede ser el número de un único residuo, un sitio entre dos números de residuos adyacentes, una región que delimite una serie de números de residuos contiguos, o un sitio o región que se extienda más allá del residuo, o de la serie de residuos que se ha especificado. Deberán utilizarse varios descriptores de localización junto con un operador de localización cuando la característica corresponda a sitios o regiones de discontinuos de la secuencia (véanse los párrafos ~~6867~~ a ~~7470~~). El descriptor de localización no deberá incluir números de residuos fuera de la serie de la secuencia indicada en el elemento `INSDSeq_sequence`.

67. La sintaxis de cada tipo descriptor de localización se indica en el cuadro presentado a continuación, donde `x` e `y` son números de residuos, indicados como enteros no negativos, no superiores a la longitud de la secuencia en el elemento `INSDSeq_sequence`, y `x` es menor que `y`.

Tipo de descriptor de localización	Sintaxis	Descripción
Número único de residuo	<code>x</code>	Designa un único residuo en la secuencia presentada .
Números de residuos que limitan un tramo de secuencia	<code>x..y</code>	Designa una serie continua de residuos delimitada por un residuo de inicio y uno de fin, incluidos éstos.
Residuos antes del primero, o después del último, residuo especificado	<code><x</code> <code>>x</code> <code><x..y</code> <code>x..>y</code>	Designa una región que incluye un residuo o una serie de residuos especificada y que se extiende más allá del residuo especificado. Los símbolos <code>'<'</code> y <code>'>'</code> podrán utilizarse con un único residuo, o los números de los residuos de inicio y de fin de una serie de residuos para indicar que la característica se extiende más allá del número de residuo especificado.

Tipo de descriptor de localización	Sintaxis	Descripción
Un sitio entre dos números de residuo adyacentes	x^y	Designa un sitio entre dos residuos adyacentes, por ejemplo, el sitio de una unión endonucleolítica. Los números de posición de los residuos adyacentes están separados por el símbolo (^). Los formatos permitidos para este descriptor son x^x+1 (por ejemplo 55^56), o para nucleótidos circulares, x^1 , donde "x" es la longitud total de la molécula, es decir, 1000^1 para una molécula circular con una longitud 1000.

68. Un operador de localización es un prefijo de un descriptor de localización o de una combinación de descriptores de localización que corresponden a una característica única pero discontinua, y especifica el lugar correspondiente a la característica en la secuencia indicada, y la manera de interpretar la característica. A continuación se suministra una lista de operadores de localización con sus respectivas definiciones.

(a) Operador de localización para nucleótidos y aminoácidos:

Sintaxis de localización	Descripción
<code>join(location,location, ... location)</code>	Las localizaciones indicadas están unidas (colocadas extremo con extremo) para formar una secuencia contigua.
<code>order(location,location, ... location)</code>	Los elementos se encuentran en el orden especificado pero ninguna información permite determinar si la unión de esos elementos es razonable.

(b) Operador de localización únicamente para nucleótidos:

Sintaxis de localización	Descripción
<code>complement(location)</code>	Indica que la característica está localizada en la cadena complementaria al tramo de la secuencia especificado por el descriptor de localización, cuando se lee en el sentido de 5' a 3' o en el sentido que imite el sentido 5' a 3'.

69. Los operadores de localización de unión (join) u orden (order) requieren al menos dos descriptores de localización separados por una coma. Los descriptores de localización relativos a sitios situados entre dos residuos adyacentes, es decir x^y , no **podrán deberán** utilizarse en una localización de unión u orden. La utilización del operador de localización join implica que los residuos designados por los descriptores de localización están puestos en contacto físicamente mediante procesos biológicos (por ejemplo, los exones que contribuyen a una característica de una región de codificación).

70. El operador de localización "complement" puede utilizarse únicamente para los nucleótidos. El operador "Complement" puede utilizarse en combinación ya sea con los operadores "join" u "order" dentro de la misma localización. No deberán utilizarse combinaciones de los operadores "join" y "order" en la misma localización.

71. Los siguientes ejemplos ilustran localizaciones de características en la forma prevista en los párrafos **6564** a **7069** *supra*:

(a) localizaciones para los nucleótidos y aminoácidos:

Ejemplo de localización	Descripción
467	Designa el residuo 467 de la secuencia.
123^124	Designa un sitio entre los residuos 123 y 124.
340..565	Designa una serie continua de residuos delimitada por los residuos 340 y 565, incluidos éstos.
<1	Designa una localización de característica antes del primer residuo.
<345..500	Indica que se desconoce el punto exacto del límite inferior de una característica. La localización comienza en algún residuo anterior al 345 y continúa hasta el residuo 500, incluido éste.
<1..888	Indica que la característica comienza antes del primer residuo de la secuencia y continúa hasta el residuo 888, incluido éste.
1..>888	Indica que la característica comienza en el primer residuo de la secuencia y continúa más allá del residuo 888.
<code>join(12..78,134..202)</code>	Indica que las regiones 12 a 78 y 134 a 202 deberían unirse para formar una secuencia contigua.

(b) localizaciones únicamente para nucleótidos:

Ejemplo de localización	Descripción
<code>complement(34..126)</code>	Comienza en la base del nucleótido complementario a 126 y termina en la base del nucleótido complementario a la base del nucleótido 34 (la característica está en la cadena complementaria a la cadena presentada).
<code>complement(join(2691..4571, 4918..5163))</code>	Une las bases los nucleótidos 2691 a 4571 y 4918 a 5163, y luego complementa los segmentos unidos (la característica está en la cadena complementaria a la cadena presentada).
<code>join(complement(4918..5163), complement(2691..4571))</code>	Complementa las regiones 4918 a 5163 y 2691 a 4571, luego une los segmentos complementados (la característica está en la cadena complementaria a la cadena presentada).

72. En una instancia XML de una lista de secuencias, los caracteres "<" y ">" de un descriptor de localización deberán sustituirse por las entidades predefinidas adecuadas (véase el párrafo 41). Por ejemplo:

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

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

Calificadores de caracterización

73. Los calificadores se utilizan para suministrar información acerca de las características que complementa la información indicada por la clave de caracterización y la localización de característica. Existen tres tipos de formatos de valor para representar los diferentes tipos de información comunicada por los calificadores, a saber:

- (a) texto libre (véanse los párrafos 8685 y 8786);
- (b) vocabulario controlado o valores enumerados (por ejemplo, un número o fecha); y
- (c) secuencias.

74. La Sección 6 del Anexo I contiene una lista completa de calificadores y sus formatos de valor especificados, dado el caso, para cada clave de caracterización de nucleótidos, y la Sección 8 contiene la lista completa de calificadores para cada clave caracterización de aminoácidos.

75. Toda secuencia contemplada en el párrafo 67 que se indique como valor calificador deberá presentarse incluirse de manera clara en la lista de secuencias ~~con~~ y se le deberá asignar su propio identificador de secuencia.

Calificadores de caracterización obligatorios

76. Una clave caracterización obligatoria, es decir, "source" para las secuencias de nucleótidos y "SOURCE" para las secuencias de aminoácidos, deberá incluir dos calificadores obligatorios, "organism" y "mol_type" para las secuencias de nucleótidos y "ORGANISM" y "MOL_TYPE" para las secuencias de aminoácidos. Algunas claves de caracterización facultativas también necesitan calificadores obligatorios.

Elementos de los calificadores

77. El elemento `INSDFeature_qual`s contiene uno o más elementos `INSDQualifier`. Cada elemento `INSDQualifier` representa un calificador único y consiste en dos elementos dependientes de la siguiente manera:

Elemento	Descripción	Obligatorio/Facultativo
<code>INSDQualifier_name</code>	Nombre del calificador (véanse las Secciones 6 y 8 del Anexo I)	Obligatorio
<code>INSDQualifier_value</code>	Valor del calificador, dado el caso, en el formato especificado (véanse las Secciones 6 y 8 del el Anexo I)	Obligatorio, cuando se especifica (véanse las Secciones 6 y 8 del el Anexo I)

78. El calificador de organismo, es decir "organism" para las secuencias de nucleótidos (véase la Sección 6 del Anexo I) y "ORGANISM" para las secuencias de aminoácidos (véase la Sección 8 del Anexo I) deberá divulgar la fuente, es decir, un organismo único u origen de la secuencia ~~que se presenta~~. Las designaciones de los organismos deberían seleccionarse a partir de una base de datos taxonómica.

79. Si la secuencia existe en estado natural y el organismo fuente tiene una designación de género y especie en latín, esa designación podrá utilizarse como valor calificador. El nombre más usual en inglés podrá especificarse utilizando el calificador "note" para las secuencias de nucleótidos y el calificador "NOTE" para las secuencias de aminoácidos, pero no deberá utilizarse en el valor calificador del organismo.

80. Los siguientes ejemplos ilustran la fuente de las **secuencias presentadas** **secuencia** en la forma prevista en los párrafos 7877 y 7978 *supra*:

Ejemplo 1: Fuente de una secuencia de nucleótidos.

```
<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>
```

Ejemplo 2: Fuente de una secuencia **proteínica** **de aminoácidos**.

```
<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>
```

81. Si la secuencia existe en estado natural y existe el nombre en latín del género del organismo fuente, pero la especie no se ha especificado o identificado, el valor calificador del organismo deberá indicar el género en latín seguido por "sp.". Por ejemplo:

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

82. Si **la fuente de** la secuencia existe en estado natural, pero se desconoce la designación en latín del género y la especie del organismo, se deberá indicar "unidentified" como valor calificador del organismo, seguido por toda información taxonómica conocida en el calificador "note" para las secuencias de nucleótidos y en el calificador "NOTE" para las secuencias de aminoácidos. Por ejemplo:

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

83. Si la secuencia existe en estado natural y el organismo fuente no tiene una designación en latín de género y especie, tal como un virus, entonces deberá utilizarse cualquier otro nombre científico aceptable (por ejemplo, "Canine adenovirus tipo 2") como valor calificador del organismo. Por ejemplo:

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


84. Si la secuencia no existe en estado natural, deberá indicarse “synthetic construct” como valor calificador del organismo. Se podrá especificar información adicional sobre la manera en que se generó la secuencia utilizando el calificador “note” para las secuencias de nucleótidos y el calificador “NOTE” para las secuencias de aminoácidos. Por ejemplo:

```
<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>
```

85. El calificador “mol_type” para las secuencias de nucleótidos (véase la Sección 6 del Anexo I) y “MOL_TYPE” para las secuencias de aminoácidos (véase la Sección 8 del Anexo I) deberá divulgar el tipo de molécula representado en la secuencia. Estos calificadores son distintos del elemento `INSDSeq_moltype` descrito en el párrafo 54:

(a) Para una secuencia de nucleótidos, el valor calificador de “mol_type” deberá ser uno de los siguientes: “genomic DNA”, “genomic RNA”, “mRNA”, “tRNA”, “rRNA”, “other RNA”, “other DNA”, “transcribed RNA”, “viral cRNA”, “unassigned DNA”, o “unassigned RNA”. Si la secuencia no existe en estado natural, es decir si el valor del calificador “organism” es “synthetic construct”, el valor calificador de “mol_type” deberá ser “other RNA” o “other DNA”;

(b) Para las secuencias de aminoácidos, el valor calificador de “MOL_TYPE” es “protein”.

Texto libre

86. El texto libre es un tipo de formato de valor para ciertos calificadores (como se describe en el Anexo I), que se presenta en forma de una frase de texto descriptiva y debería figurar preferentemente en inglés.

87. La utilización del texto libre deberá limitarse a unos cuantos términos cortos indispensables para entender una característica de la secuencia. Para cada calificador, el texto libre no deberá ser superior a 1000 caracteres.

Secuencias codificadoras

88. La clave de caracterización “CDS” podrá utilizarse para identificar secuencias codificadoras, es decir secuencias de nucleótidos que corresponden a la secuencia de aminoácidos en una proteína y el codón de terminación. El `INSDFeature_location` debería identificar la localización de la característica “CDS” y deberá incluir el codón de terminación.

89. Los calificadores “transl_table” y “translation” podrán utilizarse con la clave de caracterización “CDS” (véase el Anexo I). Cuando el calificador “transl_table” no se utiliza, se asume la utilización del cuadro de códigos normalizados (véase el Cuadro 5 de la Sección 9 del Anexo I).

90. El calificador “trans_except” deberá utilizarse con la clave de caracterización “CDS” y el calificador “translation” para identificar el codón que codifica la pirrolisina o la selenocisteína.

91. Toda secuencia **proteínica de aminoácidos** codificada por la secuencia codificadora y divulgada en un calificador "translation" contemplado en el párrafo 67 deberá **ser incluida en la lista de secuencias** y tener su propio identificador de secuencia, **y presentarse en la lista de secuencias**. El identificador de secuencia asignado a la secuencia **proteínica de aminoácidos** deberá figurar como valor del calificador "protein_id" con la clave de caracterización "CDS". El calificador "ORGANISM" de la clave de caracterización "SOURCE" para la secuencia **proteínica de aminoácidos** deberá ser idéntico al de su secuencia codificadora. Por ejemplo:

```
<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>
MLVHLERTTIMFDFSSLINLPLIWGLLIAIAVLLYILMDGFDLIGIGILLPFAPSDKCRDHMISSIAPFWDGNETWLVLGGGGL
FAAFPLAYSILMPAFYIPIIIIMLLGLIVRGVSEFRFKAEKGKRRRLWDYAFHFGSLGAAFCQGMILGAFIHGVEVNGRNFSGG
QLM
        </INSDQualifier_value>
      </INSDQualifier>
      <INSDQualifier>
        <INSDQualifier_name>protein_id</INSDQualifier_name>
        <INSDQualifier_value>89</INSDQualifier_value>
      </INSDQualifier>
    </INSDFeature_qual>
  </INSDFeature>
</INSDSeq_feature-table>
```

Variantes

92. Toda secuencia principal y toda variante de esa secuencia, cada una de ellas divulgada mediante la enumeración de sus residuos y contemplada en el párrafo 67, deberá **presentarse incluirse** en la lista de secuencias **con y se le asignará** su propio identificador de secuencia.

93. **Toda variante de secuencia, divulgada como una única secuencia con las variantes de residuos alternativos enumerados en una o más posiciones, se deberá incluir en la lista de secuencias y se debería representar con una única secuencia, en la cual las variantes de residuos alternativos enumerados se representan con el símbolo de ambigüedad más restrictivo (véanse los párrafos 15 y 27).**

94. Toda variante de secuencia, divulgada únicamente por referencia a una o varias supresiones, inserciones o sustituciones en una secuencia principal en la lista de secuencias, **podrá debería presentarse incluirse** en la lista de secuencias. Cuando se indica en la lista de secuencias, tal variante de secuencia:

- (a) podrá **presentarse representarse** por anotación de la secuencia principal, si contiene una o varias variaciones en una única localización o en varias localizaciones distintas y esas variaciones aparecen de forma independiente;
- (b) debería **presentarse representarse** como una secuencia distinta **con y se le asignará** su propio identificador de secuencia, si contiene variaciones en varias localizaciones distintas y esas variaciones aparecen de forma interdependiente; y
- (c) deberá **presentarse representarse** como una secuencia distinta **con y se le asignará** su propio identificador de secuencia, si contiene una secuencia que ha sido insertada o sustituida que contiene más de 1000 residuos (véase el párrafo 87).

95. El cuadro presentado a continuación indica la utilización adecuada de las claves y calificadores de caracterización para las variantes de ácidos nucleicos y aminoácidos:

Tipo de secuencia	Clave de caracterización	Calificador	Utilización
Ácido nucleico	variation	replace o note	Mutaciones y polimorfismos que existen en estado natural, por ejemplo, los alelos o los polimorfismos de longitud de los fragmentos de restricción.
Ácido nucleico	misc_difference	replace o note	Variabilidad introducida artificialmente, por ejemplo, mediante manipulación genética o síntesis química.
Aminoácido	VAR_SEQ	NOTE	Variante producida por empalme diferencial, utilización de promotor alternativo, iniciación alternativa y o desplazamiento del marco ribosomal.
Aminoácido	VARIANT	NOTE	Todo tipo de variante para el cual VAR_SEQ no es aplicable.

96. La anotación de una secuencia principal para una variante específica deberá contener una clave de caracterización y un calificador, como se indica en el cuadro *supra*, y la localización de la característica. El valor del calificador "replace" deberá ser exclusivamente un único nucleótido alternativo o una secuencia de nucleótidos en la cual se utilicen los símbolos que se indican en el Cuadro 1 de la Sección 1. El valor del calificador "note" o "NOTE" podrá ser una lista de variantes de residuos alternativos. En particular, se deberá indicar una lista de aminoácidos alternativos como valor del calificador "NOTE" cuando se use "X" en una secuencia, pero represente un subgrupo de "uno de los símbolos 'A', 'R', 'N', 'D', 'C', 'Q', 'E', 'G', 'H', 'I', 'L', 'K', 'M', 'F', 'P', 'O', 'S', 'U', 'T', 'W', 'Y', o 'V'". Toda supresión deberá representarse mediante un valor calificador vacío para el calificador "replace" o una indicación en el "note" o "NOTE" de que se podrá suprimir el residuo. Todo residuo insertado o sustituido deberá indicarse en el calificador "replace", "note" o "NOTE". El valor de los calificadores "replace", "note" y "NOTE" es un texto libre que no deberá ser superior a 1000 caracteres, tal como se prevé en el párrafo 8786. Para las secuencias contempladas en el párrafo 6 que se presentan como inserción o sustitución en un valor calificador, véase el párrafo 9798. El valor calificador puede contener una lista de los residuos alternativos que pueden insertarse o sustituirse.

97. Los símbolos descritos en el Anexo I (véanse los Cuadros 1 a 4 de las Secciones 1 a 4, respectivamente) deberían utilizarse para representar las variantes de residuos, dado el caso. Para el calificador "note" o "NOTE", cuando la variante de residuo es un residuo modificado que no se describe en los Cuadros 2 o 4 del Anexo I, deberá indicarse el nombre completo no abreviado del residuo modificado como valor calificador. Los residuos modificados deberán ser descritos más ampliamente en el cuadro de características según se indica en el párrafo 17 o 30.

98. Los siguientes ejemplos ilustran la presentación de variantes en la forma prevista en los párrafos 9293 a 9596 *supra*:

Ejemplo 1: Clave de caracterización "variation misc difference" para una variante de residuos alternativos de nucleótidos enumerados sustitución en una secuencia de nucleótidos. La "n" en la posición 53 de la secuencia puede ser uno de los cinco nucleótidos alternativos.

~~Una citosina sustituye al nucleótido definido en la posición 413 de la secuencia.~~

```

<INSDFeature>
  <INSDFeature_key>variation misc difference</INSDFeature_key>
  <INSDFeature_location>413-53</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>replace note</INSDQualifier_name>
      <INSDQualifier_value>ew, cmmn5s2u, 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>cmm5s2u, mam5u, mcm5s2u, or  
p</INSDQualifier_value>  
</INSDQualifier>  
</INSDFeature_qual>  
</INSDFeature>
```

Ejemplo 2: Clave de caracterización “misc_difference” para una supresión en una secuencia de nucleótidos.

Se ha suprimido el nucleótido en la posición 413.

```
<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>
```

Ejemplo 3: Clave de caracterización “misc_difference” para una inserción en una secuencia de nucleótidos.

La secuencia “atgccaaatat” se ha insertado entre las posiciones 100 y 101 de la secuencia principal.

```
<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>
```

Ejemplo 4: Clave de caracterización “**VARIANT**” **variation** para una sustitución en una secuencia de aminoácidos nucleótidos.

El aminoácido definido en la posición 100 de la secuencia puede sustituirse por I, A, F, Y, aIle, MeIle, o Nle. La citosina sustituye al nucleótido indicado en la posición 413 de la secuencia.

```
<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>
```

Ejemplo 5: Clave de caracterización “VARIANT” para una sustitución en una secuencia de aminoácidos.

El aminoácido definido en la posición 100 de la secuencia puede sustituirse por cualquier aminoácido excepto por Lys, Arg o His, I, A, F, Y, aIle, MeIle o 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>
```

Ejemplo 6: Clave de caracterización "VARIANT" para una sustitución en una secuencia de aminoácidos.

El aminoácido definido en la posición 100 de la secuencia puede sustituirse por Example 5: Feature key "VARIANT" for a substitution in an amino acid sequence. The amino acid given in position 100 of the sequence can be replaced by cualquier aminoácido, excepto por Lys, Arg o His.

99. Toda secuencia contemplada en el párrafo 67 que se indique como una inserción o sustitución en un valor calificador para una anotación de secuencia principal también deberá **presentarse** **incluirse** en la lista de secuencias **con** **se le asignará** su propio identificador de secuencia.

[Sigue el Anexo I de la Norma ST.26]

ST.26 - ANNEX I

CONTROLLED VOCABULARY

Version ~~1.01.1~~

*Proposal presented by the SEQL Task Force for consideration and approval at the CWS/5
Adopted by the Committee on WIPO Standards (CWS)
at its reconvened fourth session on March 24, 2016*

Final Draft

TABLE OF CONTENTS

SECTION 1: LIST OF NUCLEOTIDES.....	27
SECTION 2: LIST OF MODIFIED NUCLEOTIDES	27
SECTION 3: LIST OF AMINO ACIDS.....	29
SECTION 4: LIST OF MODIFIED AND UNUSUAL AMINO ACIDS.....	30
SECTION 5: FEATURE KEYS FOR NUCLEIC ACID SEQUENCES.....	31
SECTION 6: DESCRIPTION OF QUALIFIERS FOR NUCLEIC ACID SEQUENCES	49
SECTION 7: FEATURE KEYS FOR AMINO ACID SEQUENCES.....	69
SECTION 8: QUALIFIERS FOR AMINO ACID SEQUENCES.....	75
SECTION 9: GENETIC CODE TABLES.....	76

SECTION 1: LIST OF NUCLEOTIDES

The nucleotide base codes to be used in sequence listings are presented in Table 1. The symbol “t” will be construed as thymine in DNA and uracil in RNA when it is used with no further description. Where an ambiguity symbol (representing two or more bases in the alternative) is appropriate, the most restrictive symbol should be used. For example, if a base in a given position could be “a or g,” then “r” should be used, rather than “n”. The symbol “n” will be construed as “a or c or g or t/u” when it is used with no further description.

Table 1: List of nucleotides

Symbol	Nucleotide
a	adenine
c	cytosine
g	guanine
t	thymine in DNA/uracil in RNA (t/u)
m	a or c
r	a or g
w	a or t/u
s	c or g
y	c or t/u
k	g or t/u
v	a or c or g; not t/u
h	a or c or t/u; not g
d	a or g or t/u; not c
b	c or g or t/u; not a
n	a or c or g or t/u; “unknown” or “other”

SECTION 2: LIST OF MODIFIED NUCLEOTIDES

The abbreviations listed in Table 2 are the only permitted values for the mod_base qualifier. Where a specific modified nucleotide is not present in the table below, then the abbreviation “OTHER” must be used as its value. If the abbreviation is “OTHER,” then the complete unabbreviated name of the modified base must be provided in a note qualifier. The abbreviations provided in Table 2 must not be used in the sequence itself.

Table 2: List of modified nucleotides

Abbreviation	Modified Nucleotide
ac4c	4-acetylcytidine
chm5u	5-(carboxyhydroxymethyl)uridine
cm	2'-O-methylcytidine
cmnm5s2u	5-carboxymethylaminomethyl-2-thiouridine
cmnm5u	5-carboxymethylaminomethyluridine
edhu	dihydrouridine
fm	2'-O-methylpseudouridine
gal q	beta-D-galactosylguanosine galactosylqueuosine
gm	2'-O-methylguanosine
i	inosine
i6a	N6-isopentenyladenosine
m1a	1-methyladenosine
m1f	1-methylpseudouridine
m1g	1-methylguanosine
m1i	1-methylinosine
m22g	2,2-dimethylguanosine
m2a	2-methyladenosine
m2g	2-methylguanosine
m3c	3-methylcytidine
m4c	N4-methylcytosine
m5c	5-methylcytidine
m6a	N6-methyladenosine
m7g	7-methylguanosine

Abbreviation	Modified Nucleotide
mam5u	5-methylaminomethyluridine
mam5s2u	5-methoxyaminomethylmethylaminomethyl-2-thiouridine
man q	beta-D-mannosylqueosine
mcm5s2u	5-methoxycarbonylmethyl-2-thiouridine
mcm5u	5-methoxycarbonylmethyluridine
mo5u	5-methoxyuridine
ms2i6a	2-methylthio-N6-isopentenyladenosine
ms2t6a	N-((9-beta-D-ribofuranosyl-2-methylthiopurine-6-
mt6a	N-((9-beta-D-ribofuranosylpurine-6-yl)N-methyl-carbamoyl)threonine
mv	uridine-5-oxoacetic acid-methylester
o5u	uridine-5-oxoacetic acid (v)
osyw	wybutoxosine
p	pseudouridine
q	queosine
s2c	2-thiocytidine
s2t	5-methyl-2-thiouridine
s2u	2-thiouridine
s4u	4-thiouridine
m5u	5-methyluridine
t6a	N-((9-beta-D-ribofuranosylpurine-6-yl)carbamoyl)threonine
tm	2'-O-methyl-5-methyluridine
um	2'-O-methyluridine
yw	wybutosine
x	3-(3-amino-3-carboxypropyl)uridine, (acp3)u
OTHER	(requires note qualifier)

SECTION 3: LIST OF AMINO ACIDS

The amino acid codes to be used in sequence listings are presented in Table 3. Where an ambiguity symbol (representing two or more amino acids in the alternative) is appropriate, the most restrictive symbol should be used. For example, if an amino acid in a given position could be aspartic acid or asparagine, the symbol "B" should be used, rather than "X". The symbol "X" will be construed as any one of "A", "R", "N", "D", "C", "Q", "E", "G", "H", "I", "L", "K", "M", "F", "P", "O", "S", "U", "T", "W", "Y", or "V", when it is used with no further description.

Table 3: List of amino acids

Symbol	Amino acid
A	Alanine
R	Arginine
N	Asparagine
D	Aspartic acid (Aspartate)
C	Cysteine
Q	Glutamine
E	Glutamic acid (Glutamate)
G	Glycine
H	Histidine
I	Isoleucine
L	Leucine
K	Lysine
M	Methionine
F	Phenylalanine
P	Proline
O	Pyrrolysine
S	Serine
U	Selenocysteine
T	Threonine
W	Tryptophan
Y	Tyrosine
V	Valine
B	Aspartic acid or Asparagine
Z	Glutamine or Glutamic acid
J	Leucine or Isoleucine
X	unknown or other A or R or N or D or C or Q or E or G or H or I or L or K or M or F or P or O or S or U or T or W or Y or V; "unknown" or "other"

SECTION 4: LIST OF MODIFIED AND UNUSUAL AMINO ACIDS

Table 4 lists the only permitted abbreviations for a modified or unusual amino acid in the mandatory qualifier "NOTE" for feature keys "MOD_RES" or "SITE". The value for the qualifier "NOTE" must be either an abbreviation from this table, where appropriate, or the complete, unabbreviated name of the modified amino acid. The abbreviations (or full names) provided in this table must not be used in the sequence itself.

Table 4: List of modified and unusual amino acids

Abbreviation	Modified or Unusual Amino acid
Aad	2-Aminoadipic acid
bAad	3-Aminoadipic acid
bAla	beta-Alanine, beta-Aminopropionic acid
Abu	2-Aminobutyric acid
4Abu	4-Aminobutyric acid, piperidinic acid
Acp	6-Aminocaproic acid
Ahe	2-Aminoheptanoic acid
Aib	2-Aminoisobutyric acid
bAib	3-Aminoisobutyric acid
Apm	2-Aminopimelic acid
Dbu	2,4-Diaminobutyric acid
Des	Desmosine
Dpm	2,2'-Diaminopimelic acid
Dpr	2,3-Diaminopropionic acid
EtGly	N-Ethylglycine
EtAsn	N-Ethylasparagine
Hyl	Hydroxylysine
aHyl	allo-Hydroxylysine
3Hyp	3-Hydroxyproline
4Hyp	4-Hydroxyproline
Ide	Isodesmosine
alle	allo-Isoleucine
MeGly	N-Methylglycine, sarcosine
Melle	N-Methylisoleucine
MeLys	6-N-Methyllysine
MeVal	N-Methylvaline
Nva	Norvaline
Nle	Norleucine
Orn	Ornithine

SECTION 5: FEATURE KEYS FOR NUCLEIC ACID SEQUENCES

This paragraph section contains the list of allowed feature keys to be used for nucleic acid nucleotide sequences, and lists mandatory and optional qualifiers. The feature keys are listed in alphabetic order. The feature keys can be used for either DNA or RNA unless otherwise indicated under "Molecule scope". Some feature keys include a 'Parent Key' designation; when a parent key is indicated in the description of a feature key, it is mandatory that the designated parent key be used. Certain Feature Keys may be appropriate for use with artificial sequences in addition to the specified "organism scope".

Feature key names must be used in the XML instance of the sequence listing exactly as they appear following "Feature key" in the descriptions below, except for the feature keys 3'UTR and 5'UTR. See "Comment" in the description for the 3'UTR and 5'UTR feature keys.

5.1. Feature Key attenuator

Definition 1) region of DNA at which regulation of termination of transcription occurs, which controls the expression of some bacterial operons;
 2) sequence segment located between the promoter and the first structural gene that causes partial termination of transcription

Optional qualifiers allele
 gene
 gene_synonym
 map
 note
 operon
 phenotype

Organism scope prokaryotes

Molecule scope DNA

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

Parent Key CDS

Organism scope eukaryotes

5.3. Feature Key CAAT_signal

Definition CAAT box; part of a conserved sequence located about 75 bp up stream of the start point of eukaryotic transcription units which may be involved in RNA polymerase binding; consensus=GG(C or T)CAATCT [1, 2]

Optional qualifiers allele
 gene
 gene_synonym
 map
 note

Organism scope eukaryotes and eukaryotic viruses

Molecule scope DNA

References [1] Efstratiadis, A. et al. Cell 21, 653-668 (1980)
 [2] Nevins, J. R. "The pathway of eukaryotic mRNA formation" Ann Rev Biochem 52, 441-466 (1983)

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 artificial_location 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 and, pseugogene 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 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
	Parent Key	CDS
	Organism scope	eukaryotes
	5.8. Feature Key	enhancer
	Definition	a cis acting sequence that increases the utilization of (some) eukaryotic promoters, and can function in either orientation and in any location (upstream or downstream) relative to the promoter
	Optional qualifiers	allele bound_moiety gene gene_synonym map note standard_name
	Organism scope	eukaryotes and eukaryotic viruses
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
5.10.	Feature Key	GC_signal
	Definition	GC box; a conserved GC rich region located upstream of the start point of eukaryotic transcription units which may occur in multiple copies or in either orientation; consensus=GGCGGG
	Optional qualifiers	allele gene gene_synonym map note
	Organism scope	eukaryotes and eukaryotic viruses

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 Organism scope eukaryotes
		Parent Key CDS
		Organism scope eukaryotes
<hr/>		
5.15.	Feature Key	LTR
	Definition	long terminal repeat, a sequence directly repeated at both ends of a defined sequence, of the sort typically found in retroviruses
	Optional qualifiers	allele function gene gene_synonym map note standard_name
<hr/>		
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
<hr/>		
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 RBS is a regulatory class qualifier with the value "ribosome binding site" must be used for describing ribosome binding sites
<hr/>		
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 (unsure , variation, or modified_base)
	Optional qualifiers	allele clone compare gene gene_synonym map note phenotype replace standard_name
	Comment	the misc_difference feature key should 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	mi_sc_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	mi_sc_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	mi_sc_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

5.22.	Feature Key	mi_sc_signal
	Definition	any region containing a signal controlling or altering gene function or expression that cannot be described by other signal keys (promoter, CAAT_signal, TATA_signal, 35_signal, 10_signal, GC_signal, RBS, polyA_signal, enhancer, attenuator, terminator, and rep_origin)
	Optional qualifiers	allele

function
gene
gene_synonym
map
note
operon
phenotype
standard_name

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 artificial_location 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 is must not be used for ribosomal and transfer RNA annotation, for which the rRNA and tRNA feature keys should 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
	Parent Key	CDS
	Organism scope	eukaryotes

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 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 should must be used for to describe origins of replication; direction qualifier has legal values RIGHT, LEFTleft, right, and BOTHboth, however only RIGHTleft and LEFTright 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
5.31.	Feature Key	polyA_signal
	Definition	recognition region necessary for endonuclease cleavage of an RNA transcript that is followed by polyadenylation; consensus=AATAAA [1]
	Optional qualifiers	allele gene gene_synonym map note
	Organism scope	eukaryotes and eukaryotic viruses
	References	[1] Proudfoot, N. and Brownlee, G. G. Nature 263, 211-214 (1976)
5.32.		

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; includes 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 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 PCR_conditions
	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; PCR components and reaction times may be stored under the PCR_conditions qualifier; 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	promoter propeptide
	Definition	region on a DNA molecule involved in RNA polymerase binding to initiate transcription
	Optional qualifiers	allele
	Definition	propeptide coding sequence; coding sequence for the domain of a proprotein that is cleaved to form the mature protein product.
		bound_moiety
		function
		gene
		gene_synonym
		map
		note
		operon
		phenotype
		product
		pseudo
		pseudogene
		standard_name
	Molecule scope	DNA
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 RBS is the regulatory feature key and regulatory class qualifier with the value "ribosome binding site" must be used for to describe ribosome binding sites
5.31.	Feature Key	RBS regulatory
	Definition	ribosome binding site
	Optional Definition	any region of a sequence that functions in the regulation of transcription, translation, replication or chromatin structure;
	Mandatory qualifiers	allele regulatory_class gene gene_synonym map note pseudo pseudogene standard_name
	References	[1] Shine, J. and Dalgarno, L. Proc Natl Acad Sci USA 71, 1342-1346 (1974) [2] Gold, L. et al. Ann Rev Microb 35, 365-403 (1981)
	Comment	in prokaryotes, known as the Shine Dalgarno sequence; is located 5 to 9 bases upstream of the initiation codon; consensus GGAGGT [1,2]

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 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 gene gene_synonym map note standard_name
	Comment	direction qualifier has valid values: RIGHT , LEFT left, right, or BOTH 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
	Parent Key	misc_signal
	Organism scope	eukaryotes

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

5.47.	Feature Key	TATA_signal
	Definition	TATA box; Goldberg Hogness box; a conserved AT rich septamer found about 25 bp before the start point of each eukaryotic RNA polymerase II transcript unit which may be involved in positioning the enzyme for correct initiation; consensus=TATA(A or T)A(A or T) [1,2]
	Optional qualifiers	allele gene gene_synonym map note
	Organism scope	eukaryotes and eukaryotic viruses
	Molecule scope	DNA
	References	[1] Efstratiadis, A. et al. Cell 21, 653-668 (1980) [2] Corden, J., et al. "Promoter sequences of eukaryotic protein encoding genes" Science 209, 1406-1414 (1980)

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

5.40.	Feature Key	terminator
	Definition	sequence of DNA located either at the end of the transcript that causes RNA polymerase to terminate transcription
	Optional qualifiers	allele gene gene_synonym map note operon standard_name
	Molecule scope	DNA

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	author is unsure of exact sequence in this region
	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
	Parent Key	CDS
	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
	Parent Key	CDS
	Organism scope	eukaryotes
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) should must be described with the misc_difference feature; use the replace qualifier to annotate a deletion, insertion, or substitution
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 "'" in the value of an element. Thus "5' UTR" must be represented as "5'UTR" in the XML file, i.e., <INSDFeature_key>5'UTR</INSDFeature_key>.

5.59. Feature Key 10_signal

Definition Pribnow box; a conserved region about 10 bp upstream of the start point of bacterial transcription units which may be involved in binding RNA polymerase; consensus=TATAaT [1, 2, 3, 4]

Optional qualifiers allele
 gene
 gene_synonym
 map
 note
 operon
 standard_name

Organism scope prokaryotes

Molecule scope DNA

References [1] Schaller, H., Gray, C., and Hermann, K. Proc Natl Acad Sci USA 72, 737-741 (1974)
 [2] Pribnow, D. Proc Natl Acad Sci USA 72, 784-788 (1974)
 [3] Hawley, D.K. and McClure, W.R. "Compilation and analysis of Escherichia coli promoter DNA sequences" Nucl Acid Res 11, 2237-2255 (1983)
 [4] Rosenberg, M. and Court, D. "Regulatory sequences involved in the promotion and termination of RNA transcription" Ann Rev Genet 13, 319-353 (1979)

5.60. Feature Key 35_signal

Definition a conserved hexamer about 35 bp upstream of the start point of bacterial transcription units; consensus=TTGACA or TGTTGACA

Optional qualifiers allele
 gene
 gene_synonym
 map
 note
 operon
 standard_name

Organism scope prokaryotes

Molecule scope DNA

References [1] Takanami, M., et al. Nature 260, 297-302 (1976)
 [2] Moran, C.P., Jr., et al. Molec Gen Genet 186, 339-346 (1982)
 [3] Maniatis, T., et al. Cell 5, 109-113 (1975)

SECTION 6: ~~DESCRIPTION OF~~ QUALIFIERS FOR NUCLEIC ~~ACID~~ SEQUENCES

This section contains the list of qualifiers to be used for features in ~~nucleic acid~~ **nucleotide** sequences. The qualifiers are listed in alphabetic order.

Where a Value format of "none" is indicated in the description of a qualifier (e.g. germline), the INSDQualifier_value element must not be used.

PLEASE NOTE: Any qualifier value provided for a qualifier with a "free text" value format may require translation for National/Regional procedures.

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 seq<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	Multiple bound_moiety qualifiers are legal on "promoter" and "enhancer" features. 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 be specified for a given source feature 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.	Qualifier	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	DD-Mmm-YYYY, Mmm-MM-DD, YYYY-MM or YYYY
	Example	<INSDQualifier_value> 21-Oct-1952 -10-21 </INSDQualifier_value> <INSDQualifier_value> Oct-1952 -10 </INSDQualifier_value> <INSDQualifier_value>1952</INSDQualifier_value>
	Comment	full date format DD-Mmm-YYYY is preferred; where day and/or month of collection is not known either "Mmm-YYYY" or "YYYY" can be used; three letter month abbreviation can be one of the following: Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec.
6.12.	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	<INSDQualifier_value>Ni pponbare</INSDQualifier_value> <INSDQualifier_value>Tenuifolius</INSDQualifier_value> <INSDQualifier_value>Candy Cane</INSDQualifier_value> <INSDQualifier_value>IR36</INSDQualifier_value>
	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	<INSDQualifier_value>fourth instar larva</INSDQualifier_value>

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	<INSDQualifier_value>LEFTleft</INSDQualifier_value>
	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. The values are case insensitive, i.e. both "RIGHT" and "right" are valid.
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 can be replaced by dash "-" to indicate uncertain assignment. Symbol "n" can 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. Sequences; 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 <u>accordingcorresponding</u> 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 <u>(NOTE: this value may require translation for National/Regional procedures)</u>
	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 <u>(NOTE: this value may require translation for National/Regional procedures)</u>
	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 <u>(NOTE: this value may require translation for National/Regional procedures)</u>
	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 should not be used to indicate that the source of the sequence is a gamete or germ cell; germline and rearranged qualifiers cannot be used in the same source feature; germline and rearranged qualifiers should 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>8q12-13q13</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, TableSection 2
	Example	<INSDQualifier_value>m5c</INSDQualifier_value> <INSDQualifier_value>OTHER</INSDQualifier_value>
	Comment	specific modified nucleotides not found in Section 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 should be described using "genomic DNA"); ribosomal RNA genes should be described using "genomic DNA"; "rRNA" should only be used if the ribosomal RNA molecule itself has been sequenced; values "other RNA" and "other DNA" should be applied to synthetic molecules, values "unassigned DNA", "unassigned RNA" should 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 lncRNA RNase_P_RNA RNase_MRP_RNA telomerase_RNA guide_RNA rasiRNA scrRNA siRNA miRNA piRNA snoRNA snRNA SRP_RNA vault_RNA Y_RNA other
	Example	<INSDQualifier_value>autocatalytically_spliced_intron </INSDQualifier_value> <INSDQualifier_value>siRNA</INSDQualifier_value>

		<pre><INSDQualifier_value>scrNA</INSDQualifier_value> <INSDQualifier_value>other</INSDQualifier_value></pre>
Comment		specific ncRNA types not yet in the ncRNA_class controlled vocabulary can 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	<INSDQualifier_value>A comment about the feature</INSDQualifier_value>
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	<INSDQualifier_value>4</INSDQualifier_value> <INSDQualifier_value>6B</INSDQualifier_value>
	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>
	Comment	valid only on Prokaryota specific features
6. 43.	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 nucl eomorph 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>nucl eomorph</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</INSDQualifier_value>
	Example	<INSDQualifier_value>fwd_name: C01P1, fwd_seq: ttgatttttggtcayccwgaagt, rev_name: C01R4, rev_seq: ccwvytardcctarraartgttg</INSDQualifier_value> <INSDQualifier_value>fwd_name: hoge1, fwd_seq: cgkgtgtatcttact, rev_name: hoge2, rev_seq: cg<i>i>gtgtatcttact</INSDQualifier_value> <INSDQualifier_value>fwd_name: C01P1, fwd_seq: ttgatttttggtcayccwgaagt, fwd_name: C01P2, fwd_seq: gatacacaggtcayccwgaagt, rev_name: C01R4, rev_seq: ccwvytardcctarraartgttg</INSDQualifier_value>
	Comment	fwd_seq and rev_seq are both mandatory; fwd_name and rev_name are both optional. Both sequences should must be presented in 5'>3' order. The sequences should must be given in the symbols from Section 1 of this Annex, except for the modified bases; those, those, 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 pseudogenisation 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	<p>TYPE</p> <p>where TYPE is one of the following controlled vocabulary terms or phrases:</p> <p>processed</p> <p>unprocessed</p> <p>unitary</p> <p>allelic</p> <p>unknown</p>
	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	<p>Definitions of TYPE values:</p> <p>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.</p> <p>unprocessed - the pseudogene has arisen from a copy of the parent gene by duplication followed by accumulation of random mutationmutations. The changes, compared to their functional homolog, include insertions, deletions, premature stop codons, frameshifts and a higher proportion of non-synonymous versus synonymous substitutions.</p> <p>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.</p> <p>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.</p> <p>unknown - the submitter does not know the method of pseudogenisationpseudogenization.</p>
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	<p>The rearranged qualifier shouldmust not be used to annotate chromosome rearrangements that are not involved in an adaptive immune response; germline and rearranged qualifiers cannotmust not be used in the same source feature; germline and rearranged qualifiers shouldmust 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)</p>

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: <u>mitotic recombination</u> <u>non allelic homologous recombination region</u> <u>chromosome breakpoint</u>
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: <u>DNase I hypersensitive site</u> <u>enhancer blocking element</u> <u>imprinting control region</u> <u>insulator</u> <u>locus control region</u> <u>matrix attachment region</u> <u>minus_35_signal</u> <u>minus_10_signal</u> <u>recoding stimulatory region</u> <u>replication regulatory region</u> <u>response element</u> <u>polyA signal sequence</u> <u>ribosome binding site</u> <u>riboswitch</u> <u>silencer</u> <u>TATA box</u> 6.55. transcriptional cis regulatory region <u>other</u>
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)] should must be used in the CDS spans 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	organization structure and distribution of repeated sequence
	Value format	One of the following controlled vocabulary terms or phrases : tandem direct inverted flanking terminal 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 inverted </INSDQualifier_value> <INSDQualifier_value> long terminal repeat </INSDQualifier_value>
	Comment	the values are case insensitive, i.e. both "INVERTED" and "inverted" are valid; 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 which occurs as part of as set (normally a part) organized repeat pair occurring in the 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 (range) 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: S1a</INSDQualifier_value> <INSDQualifier_value>satellite: alpha</INSDQualifier_value> <INSDQualifier_value>satellite: gamma 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>0157: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	<INSDQualifier_value>female</INSDQualifier_value> <INSDQualifier_value>male</INSDQualifier_value> <INSDQualifier_value>hermaphrodite</INSDQualifier_value> <INSDQualifier_value>unisexual</INSDQualifier_value> <INSDQualifier_value>bisexual</INSDQualifier_value> <INSDQualifier_value>asexual</INSDQualifier_value> <INSDQualifier_value>monoecious</INSDQualifier_value> [or monecious] <INSDQualifier_value>dioecious</INSDQualifier_value> [or diocious]
	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	<INSDQualifier_value>dotted</INSDQualifier_value>
	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	<INSDQualifier_value>BALB/c</INSDQualifier_value>
	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	<INSDQualifier_value>lambd a-hIL7. 20g</INSDQualifier_value>
	Comment	a source feature must not contain more than one sub_clone should be specified for a given source feature qualifier; to indicate that the sequence was obtained from multiple sub_clones, multiple source features should be given 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	It must be accompanied by a strain qualifier in a source feature; if the parental strain is not given, this 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> <INSDQualifier_value>(pos: X2222: 15. . 17, aa: Ala) </INSDQualifier_value>
	Comment	if the amino acid is not one of the specific amino acids listed in Section 3 of this Annex, 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 code 'Sec' (one letter code 'U' 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 should 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 Section 3 of this Annex, "X" is to be used for AA exceptions.
	Example	<INSDQualifier_value>MASTFPPWYRGCASTPSLKGLIMCTW</INSDQualifier_value>
	Comment	to be used with CDS feature only; must be accompanied by protein_id qualifier when the translation product contains four or more specifically defined amino acids; see transl_table for definition and location of genetic code Tables; only one of the qualifiers translation, pseudo and pseudogene are permitted to further annotate a CDS feature.
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>

SECTION 7: FEATURE KEYS FOR AMINO ACID SEQUENCES

This section contains the list of allowed feature keys to be used for amino acid sequences. The feature keys are listed in alphabetic order.

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 residue of the active site should must be annotated separately with the ACT_SITE feature key. The corresponding amino acid residue number should 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); N-linked (GlcNAc...); and partial : O-linked (GlcAra...).
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 TableSection 4 of this Annex, the abbreviation may be used in place of the 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 Section 4 of this Annex, Table 4 , 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 amino acid 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"

SECTION 8: QUALIFIERS FOR AMINO ACID SEQUENCES

This section contains the list of allowed qualifiers to be used for amino acid sequences.

PLEASE NOTE: Any qualifier value provided for a qualifier with a "free text" value format may require translation for National/Regional procedures.

8.1.	Qualifier	MOL_TYPE
	Definition	In vivo molecule type of sequence
	Value format	protein
	Example	<INSDQualifier_value>prote in</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.

ST.26 - ANNEX II

DOCUMENT TYPE DEFINITION FOR SEQUENCE LISTING (DTD)

Final Draft

Proposal presented by the SEQL Task Force for consideration and adoption at the CWS/5

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

This entity may be identified by the PUBLIC identifier:

```
*****
****
```

```
PUBLIC "-//WIPO//DTD SEQUENCE LISTING 1.01//EN" "ST26SequenceListing_V1_01.dtd"
```

```
*****
****
```

```
* PUBLIC DTD URL
```

```
* http://www.wipo.int/standards/DTD/ST26SequenceListing_V1_01.dtd
```

```
*****
```

WIPO Standard ST.26, version 1.0, Recommended Standard for the presentation of nucleotide and amino acid sequence listings using XML (eXtensible Markup Language), adopted by the Committee on WIPO Standards (CWS) at its reconvened fourth session on March 24, 2016

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.
For details, please see ApplicationIdentification.
-->
<!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) of the ST.26 main body
document.
-->
<!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 International Bureau of WIPO.
-->
<!ELEMENT IPOfficeCode (#PCDATA) >

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

<!--FilingDate
The date of filing of the patent application for which the sequence listing is submitted
in ST.2 format "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. For details, please see
paragraphs 7 (a) and 11 of WIPO Standard ST.2.
-->
<!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 v1.45 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) >
```

[Annex VI to ST.26 follows]

ST.26 - ANNEX VI

GUIDANCE DOCUMENT

Final Draft

Proposal presented by the SEQL Task Force for consideration and adoption at the CWS/5

Introduction

This Standard indicates as one of its purposes, to “allow applicants to draw up a single sequence listing in a patent application acceptable for the purposes of both international and national or regional procedures.” The purpose of this Guidance Document is to ensure that all applicants and Intellectual Property Offices (IPOs) understand and agree on the requirements for inclusion and representation of sequence disclosures, such that this purpose is realized.

This guidance document consists of this introduction, an example index, examples of sequence disclosures, and an appendix containing a sequence listing in XML with sequences from the examples. This introduction explains certain concepts and terminology used in the remainder of this document. The examples illustrate the requirements of specific paragraphs of the standard and each example has been designated with the most relevant paragraph number. Some examples further illustrate other paragraphs and appropriate cross-references are indicated at the end of each example. The index provides page numbers for the examples and any indicated cross-references. Each sequence in an example that either must or may be included in a sequence listing has been assigned a sequence identification number (SEQ ID NO) and appears in XML format in the [Appendix](#) to this document.

For each example, any explanatory information presented with a sequence is intended to be considered as the entirety of the disclosure concerning that sequence. The given answers take into account only the information explicitly presented in the example.

The guidance provided in this document is directed to the preparation of a sequence listing for provision **on the filing date** of a patent application. Preparation of a sequence listing for provision **subsequent to the filing date** of a patent application must take into account whether the information provided could be considered by an IPO to add subject matter to the original disclosure. Therefore, it is possible that the guidance provided in this document may not be applicable to a sequence listing provided subsequent to the filing date of a patent application.

Preparation of a sequence listing

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

1. Does ST.26 paragraph 7 require inclusion of a particular disclosed sequence?
2. If inclusion of a particular disclosed sequence is not required, is inclusion of that sequence permitted by ST.26?
3. If inclusion of a particular disclosed sequence is required or permitted by ST.26, how should that sequence be represented in the sequence listing?

Regarding the first question, ST.26 paragraph 7 (with certain restrictions) requires inclusion of a sequence disclosed in a patent application by **enumeration of its residues**, where the sequence contains ten or more **specifically defined** nucleotides or four or more **specifically defined** amino acids.

Regarding the second question, ST.26 paragraph 8 prohibits inclusion of any sequences having fewer than ten **specifically defined** nucleotides or four **specifically defined** amino acids.

A clear understanding of “enumeration of its residues” and “specifically defined” is necessary to answer these two questions.

Regarding the third question, this document provides sequence disclosures which exemplify a variety of scenarios together with a complete discussion of the preferred means of representation of each sequence, or where a sequence contains multiple variations - the “**most encompassing sequence**”, in accordance with this Standard. Since it is impossible to address every possible unusual sequence scenario, this guidance document attempts to set forth the reasoning behind the approach to each example and the manner in which ST.26 provisions are applied, such that the same reasoning can be applied to other sequence scenarios not exemplified.

“Enumeration of its residues”

ST.26 paragraph 3(c) defines “**enumeration of its residues**” as disclosure of a sequence in a patent application by listing, in order, each residue of the sequence, wherein (i) the residue is represented by a name, abbreviation, symbol, or structure; or (ii) multiple residues are represented by a shorthand formula. A sequence should be disclosed in a patent application by “enumeration of its residues” using **conventional symbols**, which are the nucleotide symbols set forth in Section 1, Table 1 of ST.26 Annex 1 (i.e. the lower case symbols or their upper case equivalents¹) and the amino acid symbols set forth in Section 3, Table 3 of ST.26 Annex 1 (i.e. the upper case symbols or their lower case equivalents¹). Symbols other than those set forth in these tables are “**nonconventional**”.

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

Where a conventional symbol or abbreviation is used, the explanation of the sequence in the disclosure must still be consulted to confirm that the symbol is used in a conventional manner. Otherwise, if the symbol is used in a nonconventional manner, the explanation is necessary to determine whether ST.26 paragraph 7 requires inclusion in the sequence listing or whether paragraph 8 prohibits inclusion.

Where a nonconventional symbol or abbreviation is disclosed as equivalent to a conventional symbol or abbreviation (e.g., “Z₁” means “A”), or to a specific sequence of conventional symbols (e.g., “Z₁” means “agga”), then the sequence is interpreted as though it were disclosed using the equivalent conventional symbol(s) or abbreviation(s), to determine whether ST.26 paragraph 7 requires inclusion in the sequence listing or whether paragraph 8 prohibits inclusion. Where a nonconventional nucleotide symbol is used as an ambiguity symbol (e.g., X₁ = inosine or pseudouridine), but is not equivalent to one of the conventional ambiguity symbols in Section 1, Table 1 (i.e., “m”, “r”, “w”, “s”, “y”, “k”, “v”, “h”, “d”, “b”, or “n”), then the residue is interpreted as an “n” residue to determine whether ST.26 Paragraph 7 requires inclusion of the sequence in the sequence listing or whether ST.26 Paragraph 8 prohibits inclusion. Similarly, where a nonconventional amino acid symbol is used as an ambiguity symbol (e.g., “Z₁” means “A”, “G”, “S” or “T”), but is not equivalent to one of the conventional ambiguity symbols in Section 3, Table 3 (i.e., B, Z, J, or X), then the residue is interpreted as an “X” residue to determine whether ST.26 paragraph 7 requires inclusion of the sequence in the sequence listing or whether ST.26 paragraph 8 prohibits inclusion.

“Specifically defined”

ST.26 paragraph 3(k) defines “**specifically defined**” as any nucleotide other than those represented by the symbol “n” and any amino acid other than those represented by the symbol “X”, listed in Annex I, wherein “n” and “X” are used in a conventional manner as described in Section 1, Table 1 (i.e., “a or c or g or t/u; ‘unknown’ or ‘other’”) and Section 3, Table 3 (i.e., A or R or N or D or C or Q or E or G or H or I or L or K or M or F or P or O or S or U or T or W or Y or V, ‘unknown’ or ‘other’”), respectively. The discussion above concerning conventional symbols or nonconventional symbols or abbreviations and their use in a conventional or nonconventional manner will be taken into account to determine whether a nucleotide or an amino acid is “specifically defined”.

“Most encompassing sequence”

Where a sequence that meets the requirements of paragraph 7 is disclosed by enumeration of its residues only once in an application, but is described differently in multiple embodiments, e.g. in one embodiment “X” in one or more locations could be any amino acid, but in further embodiments, “X” could be only a limited number of amino acids, ST.26 requires inclusion in a sequence listing of only the single sequence that has been enumerated by its residues. As per paragraphs 15 and 27, where such a sequence contains multiple “n” or “X” ambiguity symbols, “n” or “X” is construed to represent any nucleotide or amino acid, respectively, in the absence of further annotation. Consequently, the single sequence required to be included is the most encompassing sequence disclosed. The **most encompassing sequence** is the single sequence having variant residues which are represented by the most restrictive ambiguity symbols that include the most disclosed embodiments. However, inclusion of additional specific sequences is *strongly* encouraged where practical, e.g. which represent additional embodiments that are a key part of the invention. Inclusion of the additional sequences allows for a more thorough search and provides public notice of the subject matter for which a patent is sought.

¹ NOTE: While an application disclosure may represent nucleotides or amino acids with either lower case or upper case symbols, for a sequence included in a sequence listing, only lower case letters must be used for representation of a nucleotide sequence (see ST.26 paragraph 13) and only upper case letters must be used for representation of an amino acid sequence (see ST.26 paragraph 26).

Proper Usage of the Ambiguity Symbol “n” in a Sequence Listing

The symbol “n”

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

Proper Usage of the Ambiguity Symbol “X” in a Sequence Listing

The symbol “X”

- a. must not be used to represent anything other than a single amino acid;
- b. will be construed as any one of “A”, “R”, “N”, “D”, “C”, “Q”, “E”, “G”, “H”, “I”, “L”, “K”, “M”, “F”, “P”, “O”, “S”, “U”, “T”, “W”, “Y”, or “V”, except where it is used with a further description;
- c. should be used to represent any of the following amino acids together with a further description:
 - i. modified amino acid, e.g., natural, synthetic, or non-naturally occurring, that cannot otherwise be represented by any other symbol in Annex I (see Section 3, Table 3);
 - ii. “unknown” amino acid, i.e., not determined, not disclosed, or unsure; or
- d. may be used to represent a sequence variant, i.e., alternatives, deletions, insertions, or substitutions, where “X” is the most restrictive ambiguity symbol.

Table A – Conventional Nucleotide Symbols, Abbreviations, and Names

Symbol	Abbreviation	Nucleotide Name
a		Adenine
c		Cytosine
g		Guanine
t		Thymine in DNA Uracil in RNA (t/u)
m	a or c	
r	a or g	
w	a or t/u	
s	c or g	
y	c or t/u	
k	g or t/u	
v	a or c or g; not t/u	
h	a or c or t/u; not g	
d	a or g or t/u; not c	
b	c or g or t/u; not a	
n	a or c or g or t/u; “unknown” or “other”	

Table B – Conventional Amino Acid Symbols, Abbreviations, and Names

Symbol	3-Letter Abbreviation	Amino Acid Name
A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic Acid (Aspartate)
C	Cys	Cysteine
E	Glu	Glutamic Acid (Glutamate)
Q	Gln	Glutamine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
L	Leu	Leucine
K	Lys	Lysine
M	Met	Methionine
F	Phe	Phenylalanine
P	Pro	Proline
O	Pyl	Pyrrolysine
S	Ser	Serine
U	Sec	Selenocysteine
T	Thr	Threonine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
V	Val	Valine
B	Asx	Aspartic acid or Asparagine
Z	Glx	Glutamine or Glutamic Acid
J	Xle	Leucine or Isoleucine
X	Xaa	A or R or N or D or C or Q or E or G or H or I or L or K or M or F or P or O or S or U or T or W or Y or V, "unknown" or "other"

Example Index

	Page
<u>Paragraph 3(a) – Definition of “amino acid”</u>	
Example 3(a)-1: D amino acids	96
<u>Cross-referenced examples</u>	
Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid	121
Example 30-1: Feature key “CARBOHYD”	122
<u>Paragraph 3(c) – Definition of “enumeration of its residues”</u>	
Example 3(c)-1: Enumeration of amino acids by chemical structure	97
Example 3(c)-2: Shorthand formula for an amino acid sequence	98
<u>Cross-referenced examples</u>	
Example 27-1: Shorthand formula for a nucleotide sequence	118
Example 27-3: Shorthand formula - four or more specifically defined amino acids	119
<u>Paragraph 3(f) – Definition of “modified nucleotide”</u>	
<u>Cross-referenced examples</u>	
Example 3(g)-4: Nucleic Acid Analogues	101
<u>Paragraph 3(g) – Definition of “nucleotide”</u>	
Example 3(g)-1: Nucleotide sequence interrupted by a C3 spacer	99
Example 3(g)-2: Nucleotide sequence with residue alternatives, including a C3 spacer	100
Example 3(g)-3: Abasic site	101
Example 3(g)-4: Nucleic Acid Analogues	101
<u>Cross-referenced examples</u>	
Example 11(b)-1: Double-stranded nucleotide sequence – different lengths	115
Example 14-1: The symbol “t” represents uracil in RNA	116
<u>Paragraph 3(k) – Definition of “specifically defined”</u>	
Example 3(k)-1: Nucleotide ambiguity symbols	102
Example 3(k)-2: Ambiguity symbol “n” used in both a conventional and nonconventional manner	102
Example 3(k)-3: Ambiguity symbol “n” used in a nonconventional manner	103
Example 3(k)-4: Ambiguity symbols other than “n” are “specifically defined”	104
Example 3(k)-5: Ambiguity abbreviation “Xaa” used in a nonconventional manner	104
<u>Paragraph 7 – Sequences for which inclusion in a sequence listing is required</u>	
<u>Cross-referenced examples</u>	
Example 28-1: Encoding nucleotide sequence and encoded amino acid sequence	120
Example 90-1: Amino acid sequence encoded by a coding sequence with introns	127

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

Example 7(a)-1: Branched nucleotide sequence..... 105
Example 7(a)-2: Linear nucleotide sequence having a secondary structure..... 106
Example 7(a)-3: Nucleotide ambiguity symbols used in a nonconventional manner 107
Example 7(a)-4: Nucleotide ambiguity symbols used in a nonconventional manner 108
Example 7(a)-5: Nonconventional nucleotide symbols 108
Example 7(a)-6: Nonconventional nucleotide symbols 109

Cross-referenced examples

Example 3(g)-1: Nucleotide sequence interrupted by a C3 spacer99
 Example 3(g)-2: Nucleotide sequence with residue alternatives, including a C3 spacer 100
 Example 3(g)-3: Abasic site 101
 Example 3(g)-4: Nucleic Acid Analogues 101
 Example 3(k)-1: Nucleotide ambiguity symbols..... 102
 Example 3(k)-2: Ambiguity symbol “n” used in both a conventional and nonconventional manner..... 102
 Example 3(k)-3: Ambiguity symbol “n” used in a nonconventional manner 103
 Example 3(k)-4: Ambiguity symbols other than “n” are “specifically defined” 104
 Example 11(a)-1: Double-stranded nucleotide sequence – same lengths..... 114
 Example 11(b)-1: Double-stranded nucleotide sequence – different lengths..... 115
 Example 11(b)-2: Double-stranded nucleotide sequence – no base-pairing segment..... 116
 Example 14-1: The symbol “t” represents uracil in RNA..... 116
 Example 87-1: Encoding nucleotide sequence and encoded amino acid sequence 126
 Example 91-1: Representation of enumerated variants 128
 Example 93(b)-1: Representation of individual variant sequences
 with multiple interdependent variations..... 133

Paragraph 7(b) – Amino acid sequences required in a sequence listing

Example 7(b)-1: Four or more specifically defined amino acids 110
Example 7(b)-2: Branched amino acid sequence..... 111
Example 7(b)-3: Branched amino acid sequence..... 113

Cross-referenced examples

Example 3(a)-1: D amino acids.....96
 Example 3(c)-1: Enumeration of amino acids by chemical structure97
 Example 3(c)-2: Shorthand formula for an amino acid sequence.....98
 Example 3(k)-5: Ambiguity abbreviation “Xaa” used in a nonconventional manner 104
 Example 27-1: Shorthand formula for a nucleotide sequence 118
 Example 27-3: Shorthand formula - four or more specifically defined amino acids 119

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid.....	121
Example 30-1: Feature key “CARBODHYD”	122
Example 36-1: Sequence with a region of a known number of “X” residues represented as a single sequence.....	123
Example 37-1: Sequence with regions of an unknown number of “X” residues must not be represented as a single sequence.....	125
Example 37-2: Sequence with regions of an unknown number of “X” residues must not be represented as a single sequence.....	125
Example 87-1: Encoding nucleotide sequence and encoded amino acid sequence	126
Example 91-2: Representation of enumerated variants	129
Example 91-3: Representation of a consensus sequence	130
Example 92-1: Representation of single sequence with enumerated alternative amino acids.....	131
Example 93(a)-1: Representation of a variant sequence by annotation of the primary sequence	132

Paragraph 8 – Threshold for inclusion of sequences

Cross-referenced examples

Example 3(k)-1: Nucleotide ambiguity symbols.....	102
Example 3(k)-2: Ambiguity symbol “n” used in both a conventional and nonconventional manner.....	102
Example 7(a)-1: Branched nucleotide sequence.....	105
Example 7(a)-6: Nonconventional nucleotide symbols.....	109
Example 7(b)-1: Four or more specifically defined amino acids	110
Example 14-1: The symbol “t” represents uracil in RNA.....	116
Example 37-1: Sequence with regions of an unknown number of “X” residues must not be represented as a single sequence.....	125
Example 37-2: Sequence with regions of an unknown number of “X” residues must not be represented as a single sequence.....	125
Example 92-1: Representation of single sequence with enumerated alternative amino acids.....	131

Paragraph 11 – Representation of a nucleotide sequence

Cross-referenced examples

Example 3(g)-4: Nucleic Acid Analogues	101
Example 7(a)-1: Branched nucleotide sequence.....	105

Paragraph 11(a) – Double-stranded nucleotide sequence - fully complementary

Example 11(a)-1: Double-stranded nucleotide sequence – same lengths	114
---	-----

Paragraph 11(b) – Double-stranded nucleotide sequence – not fully complementary

Example 11(b)-1: Double-stranded nucleotide sequence – different lengths.....	115
Example 11(b)-2: Double-stranded nucleotide sequence – no base-pairing segment.....	116

Paragraph 13 – Representation of nucleotides

Cross-referenced examples

Example 3(k)-2: Ambiguity symbol “n” used in both a conventional and nonconventional manner.....	102
Example 7(a)-1: Branched nucleotide sequence.....	105
Example 14-1: The symbol “t” represents uracil in RNA.....	116
Example 91-1: Representation of enumerated variants	128

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

Example 14-1: The symbol “t” represents uracil in RNA.....	116
--	-----

Paragraph 15 – The most restrictive nucleotide ambiguity symbol should be used

Cross-referenced examples

Example 3(g)-1: Nucleotide sequence interrupted by a C3 spacer	99
Example 3(g)-2: Nucleotide sequence with residue alternatives, including a C3 spacer	100
Example 3(k)-4: Ambiguity symbols other than “n” are “specifically defined”	104
Example 93(b)-1: Representation of individual variant sequences with multiple interdependent variations.....	133

Paragraph 16 – Representation of a modified amino acid

Cross-referenced examples

Example 3(g)-1: Nucleotide sequence interrupted by a C3 spacer	99
Example 3(g)-4: Nucleic Acid Analogues	101

Paragraph 17 – Annotation of a modified amino acid

Cross-referenced examples

Example 3(g)-1: Nucleotide sequence interrupted by a C3 spacer	99
Example 3(g)-3: Abasic site	101
Example 7(a)-1: Branched nucleotide sequence.....	105
Example 7(a)-6: Nonconventional nucleotide symbols.....	109

Paragraph 18 – Annotation of regions of consecutive modified nucleotides

Cross-referenced examples

Example 3(g)-4: Nucleic Acid Analogues	101
Example 11(b)-1: Double-stranded nucleotide sequence – different lengths.....	115

Paragraph 19 – Annotation of uracil in DNA or thymine in RNA

Cross-referenced examples

Example 14-1: The symbol “t” represents uracil in RNA.....	116
--	-----

Paragraph 25 – Amino acid sequence residue position number 1

Cross-referenced examples

Example 3(a)-1: D amino acids96
Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid 121

Paragraph 26 – Representation of amino acids

Cross-referenced examples

Example 36-1: Sequence with a region of a known number of “X” residues represented
as a single sequence..... 123
Example 87-1: Encoding nucleotide sequence and encoded amino acid sequence 126
Example 90-1: Amino acid sequence encoded by a coding sequence with introns 127
Example 91-2: Representation of enumerated variants 129
Example 91-3: Representation of a consensus sequence 130

Paragraph 27 – The most restrictive amino acid ambiguity symbol should be used

Example 27-1: Shorthand formula for a nucleotide sequence 118
Example 27-2: Shorthand formula - less than four specifically defined amino acids 118
Example 27-3: Shorthand formula - four or more specifically defined amino acids 119

Cross-referenced examples

Example 7(b)-1: Four or more specifically defined amino acids 110
Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid 121
Example 36-1: Sequence with a region of a known number of “X” residues represented
as a single sequence..... 123
Example 36-2: Sequence with multiple regions of a known number or range of “X” residues
represented as a single sequence..... 123
Example 36-3: Sequence with multiple regions of a known number or range of “X” residues
represented as a single sequence..... 124
Example 91-3: Representation of a consensus sequence 130
Example 92-1: Representation of single sequence with enumerated alternative amino acids..... 131
Example 93(a)-1: Representation of a variant sequence by annotation of the primary sequence 132

Paragraph 28 – Amino acid sequences separated by internal terminator symbols

Example 28-1: Encoding nucleotide sequence and encoded amino acid sequence 120

Cross-referenced examples

Example 87-1: Encoding nucleotide sequence and encoded amino acid sequence 126
Example 90-1: Amino acid sequence encoded by a coding sequence with introns..... 127

Paragraph 29 – Representation of an “other” modified amino acid

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid 121

Cross-referenced examples

Example 3(a)-1: D amino acids96

Example 30-1: Feature key “CARBODHYD” 122

Paragraph 30 – Annotation of a modified amino acids

Example 30-1: Feature key “CARBODHYD” 122

Cross-referenced examples

Example 3(a)-1: D amino acids96

Example 3(c)-1: Enumeration of amino acids by chemical structure97

Example 7(b)-2: Branched amino acid sequence.....111

Example 7(b)-3: Branched amino acid sequence.....113

Paragraph 31 – Representation of a D-amino acid

Cross-referenced examples

Example 3(a)-1: D amino acids96

Example 3(c)-1: Enumeration of amino acids by chemical structure97

Example 7(b)-2: Branched amino acid sequence.....111

Example 7(b)-3: Branched amino acid sequence.....113

Paragraph 32 – Annotation of an “unknown” amino acid

Cross-referenced examples

Example 3(c)-1: Enumeration of amino acids by chemical structure97

Paragraph 34 – Annotation of a contiguous region of “X” residues

Cross-referenced examples

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid 121

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

Example 36-1: Sequence with a region of a known number of “X” residues represented as a single sequence..... 123

Example 36-2: Sequence with multiple regions of a known number or range of “X” residues represented as a single sequence..... 123

Example 36-3: Sequence with multiple regions of a known number or range of “X” residues represented as a single sequence..... 127

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

Example 37-1: Sequence with regions of an unknown number of “X” residues must not be represented as a single sequence 125

Example 37-2: Sequence with regions of an unknown number of “X” residues must not be represented as a single sequence 125

Paragraph 54 – The element INSDSeq_moltype

Cross-referenced examples

Example 14-1: The symbol “t” represents uracil in RNA..... 116

Paragraph 57 – The element INSDSeq_sequence

Cross-referenced examples

Example 28-1: Encoding nucleotide sequence and encoded amino acid sequence 120

Example 90-1: Amino acid sequence encoded by a coding sequence with introns..... 127

Paragraph 65 – Location descriptor

Cross-referenced examples

Example 3(g)-4: Nucleic Acid Analogues 101

Paragraph 66 – Location descriptor syntax

Cross-referenced examples

Example 3(g)-4: Nucleic Acid Analogues 101

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid..... 121

Paragraph 70 – Feature locations

Cross-referenced examples

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid..... 121

Paragraph 71 – Representation of the characters “<” and “>” in a location descriptor

Cross-referenced examples

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid..... 121

Paragraph 87 – “CDS” Feature key

Example 87-1: Encoding nucleotide sequence and encoded amino acid sequence 126

Cross-referenced examples

Example 90-1: Amino acid sequence encoded by a coding sequence with introns..... 127

Paragraph 88 – The qualifiers “transl_table” and “translation”

Cross-referenced examples

Example 28-1: Encoding nucleotide sequence and encoded amino acid sequence 120

Example 87-1: Encoding nucleotide sequence and encoded amino acid sequence 126

Example 90-1: Amino acid sequence encoded by a coding sequence with introns..... 127

Paragraph 90 – Encoded amino acid sequence inclusion in a sequence listing

Example 90-1: Amino acid sequence encoded by a coding sequence with introns 127

Cross-referenced examples

Example 28-1: Encoding nucleotide sequence and encoded amino acid sequence 120

Example 87-1: Encoding nucleotide sequence and encoded amino acid sequence 126

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

Example 91-1: Representation of enumerated variants..... 128

Example 91-2: Representation of enumerated variants..... 129

Example 91-3: Representation of a consensus sequence 130

Paragraph 92 – Variant sequence disclosed as a single sequence with enumerated alternative residues

Example 92-1: Representation of single sequence with enumerated alternative amino acids 131

Paragraph 93(a) – A variant sequence disclosed only by reference to a primary sequence with multiple independent variations

Example 93(a)-1: Representation of a variant sequence by annotation of the primary sequence..... 132

Paragraph 93(b) – A variant sequence disclosed only by reference to a primary sequence with multiple interdependent variations

Example 93(b)-1: Representation of individual variant sequences with multiple interdependent variations 133

Paragraph 94 – Feature keys and qualifiers for a variant sequence

Cross-referenced examples

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid..... 121

Paragraph 95– Annotation of a variant sequence

Cross-referenced examples

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid..... 121

Example 91-3: Representation of a consensus sequence 130

Example 92-1: Representation of single sequence with enumerated alternative amino acids..... 131

Examples

Paragraph 3(a) Definition of “amino acid”

Example 3(a)-1: D amino acids

A patent application describes the following sequence:

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

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

YES

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

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

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

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

Paragraph 25 indicates that when amino acid sequences are circular in configuration, applicant must choose the amino acid in residue position number 1. Accordingly, the sequence may be represented as:

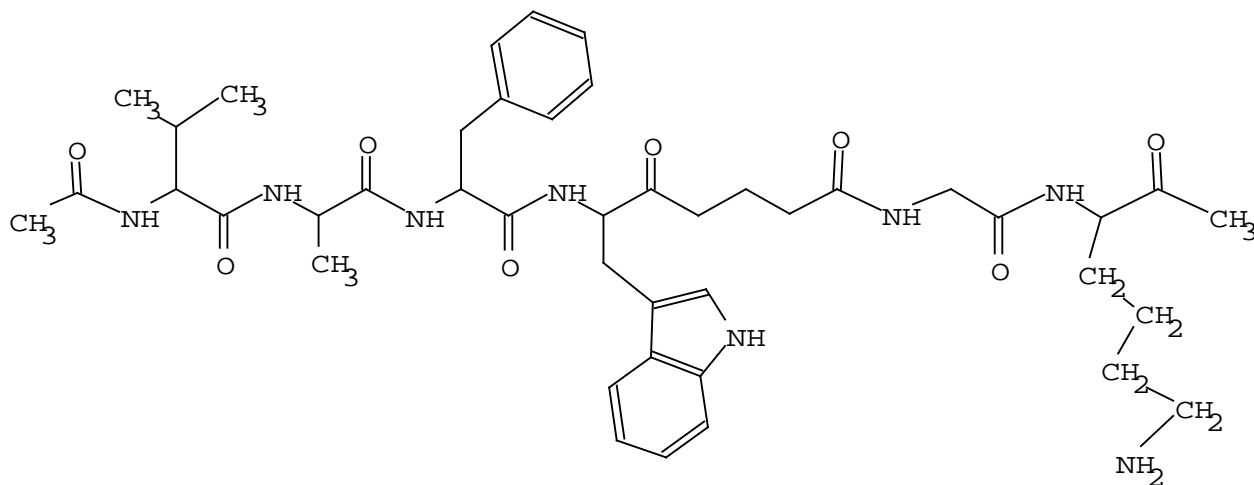
AEKXGMX (SEQ ID NO: 1)

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

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

Paragraph 3(c) – Definition of “enumeration of its residues”

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



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

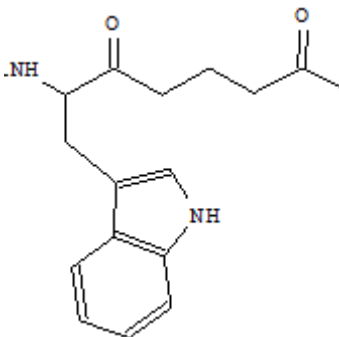
YES

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

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

The sequence may be represented as:

VAFXGK (SEQ ID NO: 2)



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

Alternatively, the sequence may be represented as:

VAFW (SEQ ID NO: 3)

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

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

Example 3(c)-2: Shorthand formula for an amino acid sequence



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

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

Yes

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

The enumerated peptide in the unexpanded formula (“n” = 1) provides four specifically defined amino acids, each of which is Gly, and the symbol “z”. Conventionally “Z” is the symbol for “glutamine or glutamic acid”; however, the example defines “z” as “any amino acid”. Under ST.26, an amino acid that is not specifically defined is represented by “X”. Based on this analysis, the enumerated peptide, i.e. GGGGX, contains four glycine residues that are enumerated and specifically defined. Thus, ST.26 paragraph 7(b) requires inclusion of the sequence in a sequence listing.

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

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

GGGGX (SEQ ID NO: 4)

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

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

Relevant ST.26 paragraph(s): Paragraph 3(c) and 7(b)

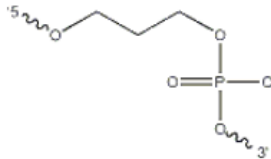
Paragraph 3(g) Definition of “nucleotide”

Example 3(g)-1: Nucleotide sequence interrupted by a C3 spacer

A patent application describes the following sequence:

atgcatgcatgcncggcatgcatgc

where n = a C3 spacer with the following structure:



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

YES

The enumerated sequence contains two segments of specifically defined nucleotides separated by a C3 spacer.

The C3 spacer is not a nucleotide according to paragraph 3(g); the conventional symbol “n” is being used in a nonconventional manner (see Introduction to this document). Consequently, each segment is a separate nucleotide sequence. Since each segment contains more than 10 specifically defined nucleotides, both must be included in a sequence listing.

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

Each segment must be included in a sequence listing as a separate sequence, each with their own sequence identification number:

atgcatgcatgc (SEQ ID NO: 5)

cggcatgcatgc (SEQ ID NO: 6)

The cytosine in each segment that is attached to the C3 spacer should be further described in a feature table using the feature key “misc_feature” and the qualifier “note”. The “note” qualifier value, which is “free text”, should indicate the presence of the spacer, which is joined to another nucleic acid.

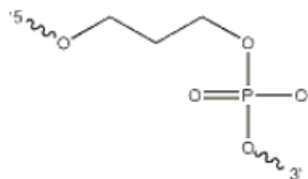
Relevant ST.26 paragraphs: Paragraphs 3(g), 7(a), and 15

Example 3(g)-2: Nucleotide sequence with residue alternatives, including a C3 spacer

A patent application describes the following sequence:

atgcatgcatgcnccggcatgcatgc

where n = c, a, g, or a C3 spacer with the following structure:



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

YES

There are 24 specifically defined residues in the enumerated sequence interrupted by the variable “n.” The explanation of the sequence in the disclosure must be consulted to determine if the “n” is used in a conventional or nonconventional manner (see Introduction to this document).

The disclosure indicates that n = c, a, g, or a C3 spacer. The “n” is a conventional symbol used in a nonconventional manner, since it is described as including a C3 spacer, which does not meet the definition of a nucleotide. The symbol “n” is also described as including “c”, “a”, or “g”; therefore, ST.26 requires inclusion of the 25 nucleotide sequence in a sequence listing. Since two segments separated by the C3 spacer are distinct sequences from the 25 nucleotide sequence, the two 12 nucleotide sequences may also be included.

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

The example indicates that “n = c, a, g, or a C3 spacer”. As discussed above, a C3 spacer is not a nucleotide. According to paragraph 15, the symbol “n” must not be used to represent anything other than a nucleotide; therefore, the symbol “n” cannot represent a C3 spacer in a sequence listing.

Paragraph 15 also states that where an ambiguity symbol is appropriate, the most restrictive symbol should be used. The symbol “v” represents “a or c or g” according to Annex I, Section 1, Table 1, which is more restrictive than “n”.

Where variable “n” in the example is c, a, or g, the single sequence enumerated by its residues that includes the most disclosed embodiments, and is therefore, the most encompassing sequence (see Introduction to this document) that must be included in a sequence listing is:

atgcatgcatgcvccggcatgcatgc (SEQ ID NO: 7)

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

Where variable “n” in the example is a C3 spacer, the sequence can be considered two separate segments of specifically defined nucleotides on either side of the variable “n”, i.e. atgcatgcatgc (SEQ ID NO: 8); and cggcatgcatgc (SEQ ID NO: 9). If essential to the disclosure or claims, these two sequences should also be included in the sequence listing, each with their own sequence identification number.

The cytosine in each segment that is attached to the C3 spacer should be further described in a feature table using the feature key “misc_feature” and the qualifier “note”. The “note” qualifier value, which is “free text”, should indicate the presence of the spacer, which is joined to another nucleic acid and identify the spacer by either its complete unabbreviated chemical name, or by its common name, e.g. C3 spacer.

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

Relevant ST.26 paragraphs: Paragraphs 3(g), 7(a), and 15

Example 3(g)-3: Abasic site

A patent application describes the following sequence:

gagcattgac-AP-taaggct

Wherein AP is an abasic site

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

YES

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

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

The sequence must be included in a sequence listing as:

gagcattgacntaaggct (SEQ ID NO: 10)

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

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

Example 3(g)-4: Nucleic Acid Analogues

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

PO₄-tagttcattgactaaggctccccattgact-OH

Wherein the left end of the sequence mimics the 5' end of a DNA sequence.

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

YES – The individual residues that comprise a GNA sequence are considered nucleotides according to ST.26 paragraph 3(g)(i)(2). Accordingly, the sequence has more than ten enumerated and "specifically defined" nucleotides and is required to be included in a sequence listing.

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

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

tagttcattgactaaggctccccattgact (SEQ ID NO: 11)

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

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

Paragraph 3(k) Definition of “specifically defined”

Example 3(k)-1: Nucleotide ambiguity symbols

5' NNG KNG KNG K 3'

N and K are IUPAC-IUB ambiguity codes

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

NO

IUPAC-IUB ambiguity codes correspond to the list of nucleotide symbols defined in Annex I, Section 1, Table 1. According to paragraph 3(k), a specifically defined nucleotide is any nucleotide other than those represented by the symbol “n” listed in Annex I. Therefore, “K” and “G” are specifically defined nucleotides and “N” is not a specifically defined nucleotide.

The enumerated sequence does not have ten or more specifically defined nucleotides and therefore is not required by ST.26 paragraph 7(a) to be included in a sequence listing.

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

NO

According to paragraph 8, “A sequence listing must not include any sequences having fewer than ten specifically defined nucleotides....” The enumerated sequence does not have ten or more specifically defined nucleotides; therefore, it must not be included in a sequence listing.

Relevant ST.26 paragraphs: Paragraphs 3(k), 7(a), 8, and 13

Example 3(k)-2: Ambiguity symbol “n” used in both a conventional and nonconventional manner

An application discloses the artificial sequence: 5'-AATGCCGGAN-3'. The disclosure further states:

- (i) in one embodiment, N is any nucleotide;
- (ii) in one embodiment, N is optional but is preferably G;
- (iii) in one embodiment, N is K;
- (iv) in one embodiment, N is C.

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

NO

The enumerated sequence contains 9 specifically defined nucleotides and an “N.” The explanation of the sequence in the disclosure must be consulted to determine if the symbol “N” is used in a conventional manner (see Introduction to this document).

Consideration of disclosed embodiments (i) through (iv) of the enumerated sequence reveals that the most encompassing embodiment of “N” is “any nucleotide”. In the most encompassing embodiment, “N” in the enumerated sequence is used in a conventional manner.

In certain embodiments “N” is described as specifically defined residues (i.e., “N is C” in part (iv)). However, only the most encompassing embodiment (i.e., “N is any nucleotide”) is considered when determining if a sequence must be included in a sequence listing. Thus, the enumerated sequence that must be evaluated is 5'-AATGCCGGAN-3'.

Based on this analysis, the enumerated sequence, i.e. AATGCCGGAN, does not contain ten specifically defined nucleotides. Therefore, ST.26 paragraph 7(a) does not require inclusion of the sequence in a sequence listing, despite the fact that “n” is also defined as specific nucleotides in some embodiments.

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

NO

The sequence "AATGCCGGAN" must not be included in a sequence listing.

However, a described alternative sequence may be included in a sequence listing if the "N" is replaced with a specifically defined nucleotide.

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

Inclusion of sequences which represent embodiments that are a key part of the invention is **strongly** encouraged. Inclusion of these sequences allows for a more thorough search and provides public notice of the subject matter for which a patent is sought.

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

aatgccggag (SEQ ID NO: 12)

aatgccggak (SEQ ID NO: 13)

aatgccggac (SEQ ID NO: 14)

If less than all three of the above sequences are included, the nucleotide that replaces the "n" should be annotated to describe the alternatives. For example, if only SEQ ID NO: 12 above is included in the sequence listing, the feature key "misc_difference" with feature location "10" should be used together with two "replace" qualifiers where the value for one would be "g" and the second would be "c".

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

Relevant ST.26 paragraphs: Paragraphs 3(k), 7(a), 8, and 13

Example 3(k)-3: Ambiguity symbol "n" used in a nonconventional manner

An application discloses the sequence: 5'-aatgttggan-3'

Wherein n is c

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

YES

According to paragraph 3(k), a "specifically defined" nucleotide is any nucleotide other than those represented by the symbol "n" listed in Annex I, Section 1, Table 1.

In this example "n" is used in a nonconventional manner to represent only "c". The disclosure does not indicate that "n" is used in the conventional manner to represent "any nucleotide". Therefore, the sequence must be interpreted as if the equivalent conventional symbol, i.e. "c", had been used in the sequence (see Introduction to this document). Accordingly, the enumerated sequence that must be considered is:

5'-aatgttggac-3'

This sequence has ten specifically defined nucleotides and is required by ST.26 paragraph 7(a) to be included in a sequence listing.

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

The sequence must be included in a sequence listing as: aatgttggac (SEQ ID NO: 15)

Relevant ST.26 paragraphs: Paragraphs 3(k) and 7(a)

Example 3(k)-4: Ambiguity symbols other than “n” are “specifically defined”

A patent application describes the following sequence:

5' NNG KNG KNG KAG VCR 3'

wherein N, K, V, and R are IUPAC-IUB ambiguity codes

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

YES

IUPAC-IUB ambiguity codes correspond to the list of nucleotide symbols defined in Annex I, Section 1, Table 1. According to paragraph 3(k), a “specifically defined” nucleotide is any nucleotide other than those represented by the symbol “n” listed in Annex I, Section 1, Table 1. Therefore, “K”, “V”, and “R” are “specifically defined” nucleotides.

The sequence has eleven enumerated and “specifically defined” nucleotides and is required by ST.26 paragraph 7(a) to be included in a sequence listing.

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

The sequence must be included in a sequence listing as:

nngkngkngkagvcr (SEQ ID NO: 16)

Relevant ST.26 paragraphs: Paragraphs **3(k)**, 7(a) and 15

Example 3(k)-5: Ambiguity abbreviation “Xaa” used in a nonconventional manner

A patent application describes the following sequence:

Xaa-Tyr-Glu-Xaa-Xaa-Xaa-Leu

Wherein Xaa in position 1 is any amino acid, Xaa in position 4 is Lys, Xaa in position 5 is Gly and Xaa in position 6 is Leucine or Isoleucine.

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

YES

The enumerated peptide in the formula provides three specifically defined amino acids in positions 2, 3 and 7. The first amino acid is represented by a conventional abbreviation, i.e., Xaa, representing any amino acid. However, the 4th, 5th and 6th amino acids are represented by a conventional abbreviation used in a nonconventional manner (see Introduction to this document). Therefore, the explanation of the sequence in the disclosure is consulted to determine the definition of “Xaa” in these positions. Since “Xaa” in positions 4-6 are indicated as a specific amino acid, the sequence must be interpreted as if the equivalent conventional abbreviations had been used in the sequence, i.e. Lys, Gly, and (Leu or Ile). Consequently, the sequence contains four or more specifically defined amino acids and must be included in a sequence listing as required by ST.26 paragraph 7(b).

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

The sequence uses a conventional abbreviation “Xaa” in a nonconventional manner. Therefore, the explanation of the sequence in the disclosure must be consulted to determine the definition of “Xaa” in positions 4, 5 and 6. The explanation defines “Xaa” as a lysine in position 4, a glycine in position 5 and a leucine or isoleucine in position 6. The conventional symbols for these amino acids are K, G, and J respectively. Therefore, the sequence should be represented as in the sequence listing as:

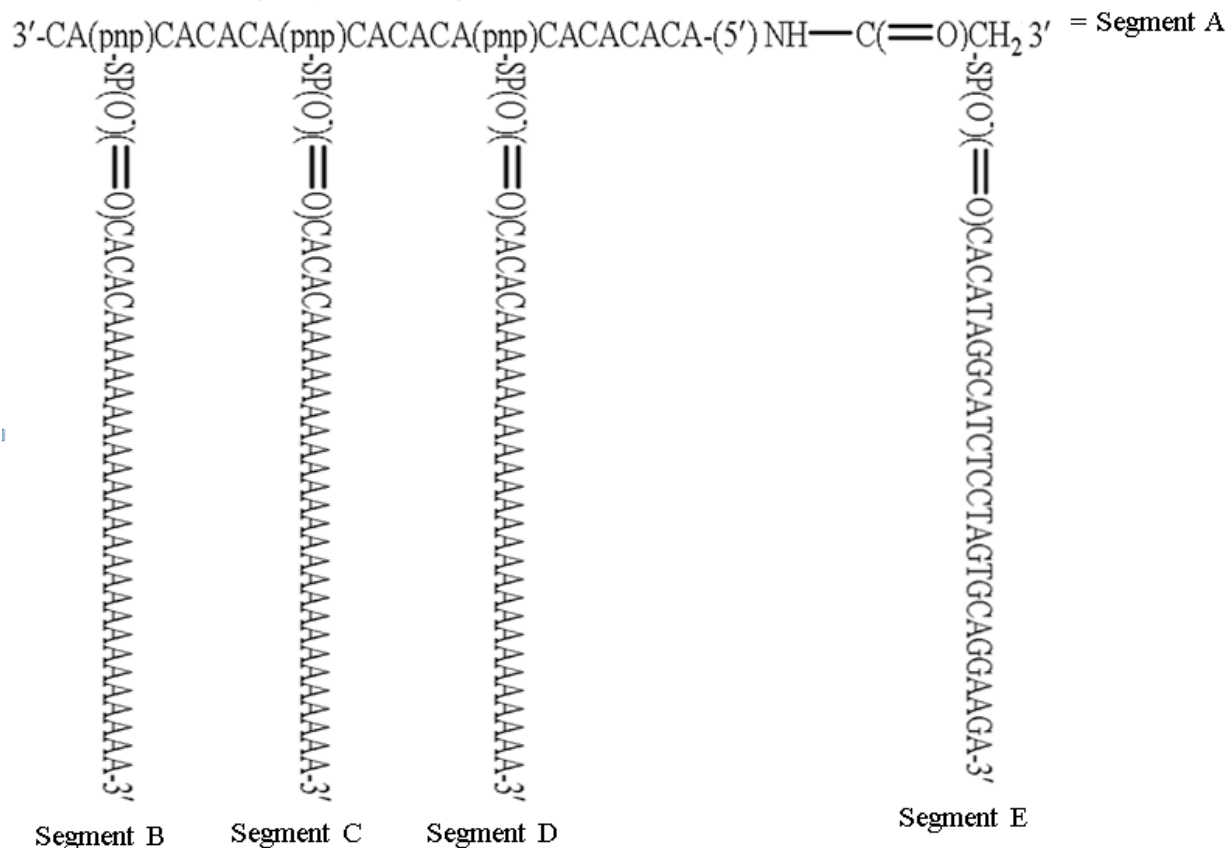
XYEKGJL (SEQ ID NO: 17)

Relevant ST.26 paragraphs: Paragraphs **3(k)**, 7(b), 26, and 27

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

Example 7(a)-1: Branched nucleotide sequence

The description discloses the following branched nucleotide sequence:



wherein "pnp" is a linkage or monomer containing an bromoacetyl amino functionality;
 3'-CA(pnp)CACACA(pnp)CACACA(pnp)CACACACA-(5')NH—C(=O)CH₂ 3' is segment A;
 SP(O)(=O)CACACAAAAAAAAAAAAAAAAAAAAAAAAA 3' is segments B, C, and D; and
 SP(O)(=O)CACATAGGCATCTCCTAGTGCAGGAAGA 3' is segment E.

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

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

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

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

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

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

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

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

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

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

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

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

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

cacacaaaaaaaaaaaaaaaaaaaaaa. (SEQ ID NO: 18)

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

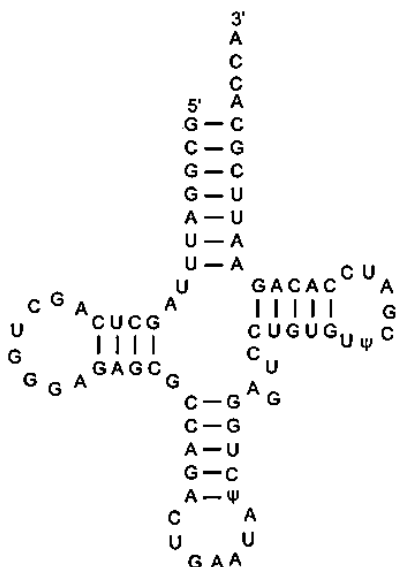
Segment E must be included in a sequence listing as a single sequence:

cacataggcatctcctagtcaggaaga. (SEQ ID NO: 19)

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

Relevant ST.26 paragraph(s): Paragraphs 7(a), 8, 11, 13, and 17

Example 7(a)-2: Linear nucleotide sequence having a secondary structure



Wherein Ψ is pseudouridine.

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

YES

The nucleotide sequence contains seventy-three enumerated and specifically defined nucleotides. Thus, the example has ten or more “specifically defined” nucleotides, and as required by ST.26 paragraph (7)(a), must be included in a sequence listing.

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

Consultation of the disclosure indicates that “ Ψ ” is equivalent to pseudouridine. The only conventional symbol that can be used to represent pseudouridine is “n”; therefore, the “ Ψ ” is a nonconventional symbol used to represent the conventional symbol “n” (see Introduction to this document). Accordingly, the sequence must be interpreted to have two “n” symbols in place of the two “ Ψ ” symbols.

The symbol “u” must not be used to represent uracil in an RNA molecule in the sequence listing. According to paragraph 14, the symbol “t” will be construed as uracil in RNA. The sequence must be included as:

gcggttagctcagctggagagcgcagactgaatanctggagtcctgtgncgatccacagaattcgcacca (SEQ ID NO: 20)

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

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

Relevant ST.26 paragraph(s): Paragraphs 7(a), 11, 13, 14, 62, 84 and Annex I, sections 2 and 5, feature key 5.43

Example 7(a)-3: Nucleotide ambiguity symbols used in a nonconventional manner

A patent application describes the following sequence:

5' GATC-MDR-MDR-MDR-MDR-GTAC 3'

The explanation of the sequence in the disclosure further indicates: “A “DR Element” consists of the sequence 5' ATCAGCCAT 3'. A mutant DR Element, or MDR, is a DR element wherein the middle 5 nucleotides, CAGCC, are mutated to TTTTT.”

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

YES

The enumerated sequence uses the symbol “MDR”. Where it is unclear if a symbol used in a sequence is intended to be a conventional symbol, i.e., a symbol set forth in Annex 1, Section 3, Table 3, or a nonconventional symbol, the explanation of the sequence in the disclosure must be consulted to make a determination (see Introduction to this document). According to Table 3, “MDR” could be interpreted as three conventional symbols (m = a or c, d = a or g or t/u, r = g or a) or as an abbreviation that is short-hand notation for some other structure.

Consultation of the disclosure indicates that an MDR element is equivalent to 5' ATTTTTTAT 3'. The letters “MDR” are considered conventional symbols used in a nonconventional manner; therefore, the sequence must be interpreted as though it were disclosed using the equivalent conventional symbols. Accordingly, the enumerated sequence that is considered for inclusion in a sequence listing is:

5' GATC ATTTTTTAT ATTTTTTAT ATTTTTTAT ATTTTTTAT GTAC 3'

The enumerated sequence has 44 specifically defined nucleotides and is required by ST.26 paragraph 7(a) to be included in a sequence listing.

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

The sequence must be included in a sequence listing as:

gatcatttttatatttttatatttttatatttttatgtac (SEQ ID NO: 21)

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

Example 7(a)-4: Nucleotide ambiguity symbols used in a nonconventional manner

A patent application describes the following sequence:

5' ATTC-N-N-N-N-GTAC 3'

The explanation of the sequence in the disclosure further indicates that "N" consists of the sequence 5' ATACGCACT 3'.

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

YES

The enumerated sequence uses the symbol "N". The explanation of the sequence in the disclosure must be consulted to determine if the "N" is used in a conventional or nonconventional manner (see Introduction to this document).

Consultation of the disclosure indicates that "N" is equivalent to 5' ATACGCACT 3'. Thus, the "N" is a conventional symbol used in a nonconventional manner. Accordingly, the sequence must be interpreted as though it were disclosed using the equivalent conventional symbols:

5' ATTC-ATACGCACT-ATACGCACT-ATACGCACT-ATACGCACT-GTAC 3'

The enumerated sequence has 44 specifically defined nucleotides and is required by ST.26 paragraph 7(a) to be included in a sequence listing.

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

The sequence must be included in a sequence listing as:

atctacatcgactatatacgactatatacgactatatacgactgtac (SEQ ID NO: 22)

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

Example 7(a)-5: Nonconventional nucleotide symbols

A patent application describes the following sequence:

5' GATC-β-β-β-β-GTAC 3'

The explanation of the sequence in the disclosure further indicates that "β" consists of the sequence 5' ATACGCACT 3'.

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

YES

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

Consultation of the disclosure indicates that "β" is equivalent to 5' ATACGCACT 3'. Thus, the "β" is a nonconventional symbol used to represent a sequence of nine specifically defined, conventional symbols. Accordingly, the sequence must be interpreted as though it were disclosed using the equivalent conventional symbols:

5' GATC-ATACGCACT-ATACGCACT-ATACGCACT-ATACGCACT-GTAC 3'

The enumerated sequence has 44 specifically defined nucleotides and is required by ST.26 paragraph 7(a) to be included in a sequence listing.

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

The sequence must be included in a sequence listing as:

gatcatatcgactatatacgactatatacgactatatacgactgtac (SEQ ID NO: 23)

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

Example 7(a)-6: Nonconventional nucleotide symbols

A patent application describes the following sequence:

5' GATC-β-β-β-GTAC 3'

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

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

NO

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

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

5' GATC-N-N-N-N-GTAC 3'

The enumerated sequence has only eight specifically defined nucleotides and is not required by ST.26 paragraph 7(a) to be included in a sequence listing.

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

NO

The enumerated sequence, 5' GATC-N-N-N-N-GTAC 3' must not be included in a sequence listing.

However, a disclosed alternative sequence may be included in a sequence listing if at least 2 of the “n” symbols are replaced by adenine, resulting in a sequence with at least 10 or more specifically defined nucleotides.

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

One possible permitted representation is:

gatcaaaagtac (SEQ ID NO: 24)

In the above example, the four adenine nucleotides that replace the β symbols should be annotated to note that these positions could be substituted with inosine or pseudouridine.

The feature key “misc_difference” should be used with a feature location 5-8 and a qualifier “note” with the value, e.g., “A nucleotide in any of positions 5-8 may be replaced with inosine or pseudouridine”. Since these alternatives are modified nucleotides, then the feature key “modified_base” together with the qualifier “mod_base” would be required. The value for the “mod_base” qualifier can be “OTHER” with a “note” qualifier and the value of “i or p”.

Other permutations are possible.

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

Relevant ST.26 paragraphs: Paragraph 7(a), 8, 13, and 17

Paragraph 7(b) – Amino Acid sequences required in a sequence listing

Example 7(b)-1: Four or more specifically defined amino acids

XXXXXXXXDXXXXXXXXXXFXXGXXXXX

Where X = any amino acid

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

YES

The enumerated peptide contains four specifically defined amino acids. The symbol “X” is used conventionally to represent the remaining amino acids as any amino acid (see Introduction to this document).

Because there are four specifically defined amino acids, i.e., Asp, Phe, Ala and Gly, ST.26 paragraph 7(b) requires that the sequence be included in a sequence listing.

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

The sequence must be represented as:

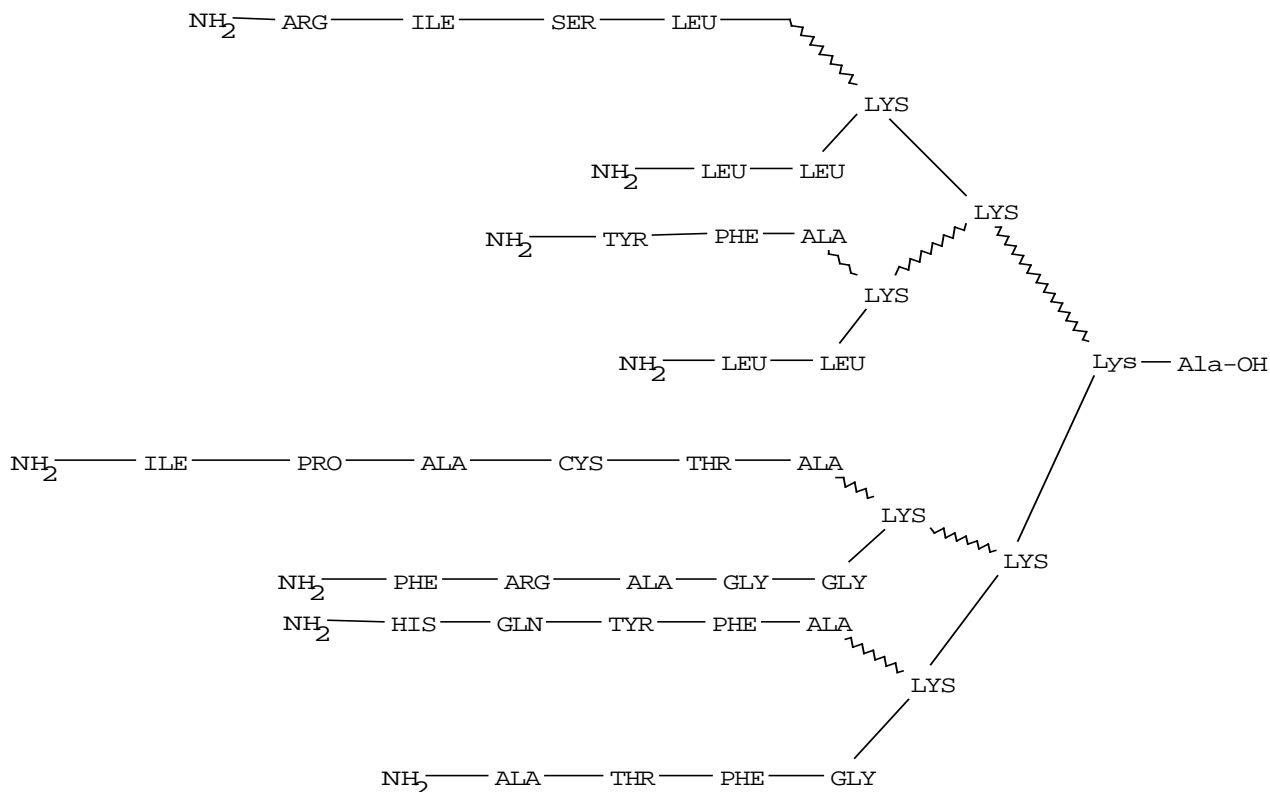
XXXXXXXXDXXXXXXXXXXFXXGXXXXX
(SEQ ID NO: 25)

Since “X” can be any amino acid, annotation of the “X” residues is not required under paragraph 27.

Relevant ST.26 paragraph(s): Paragraphs **7(b)**, 8 and 27

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

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

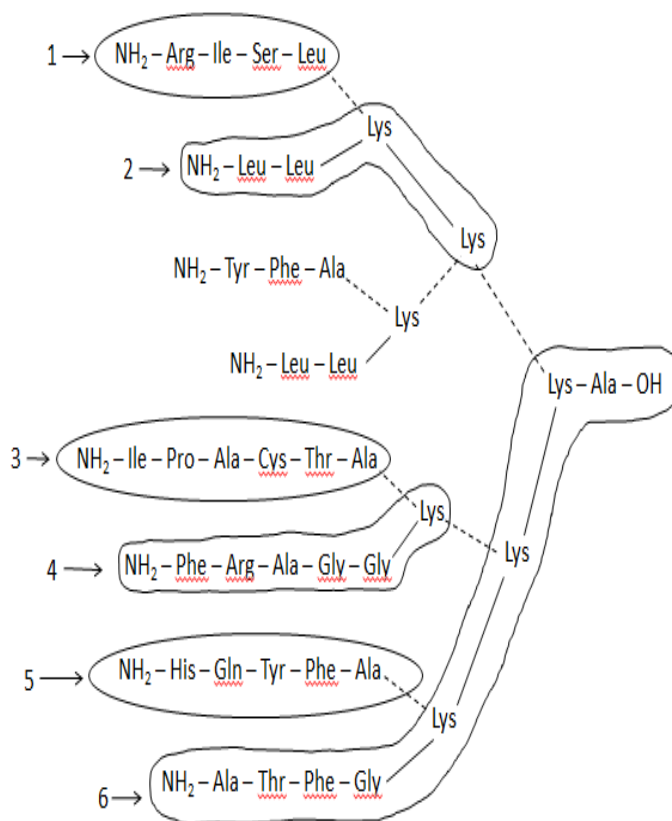


In the above branched peptide, the bonds depicted by — represent an amide linkage between the terminal amine of the Lysine and the carboxyl end of the bonded amino acid. The bonds depicted by ~ represent an amide linkage between the side chain amine of the Lysine and the carboxyl end of the bonded amino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

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

YES

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



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

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

YFA

LLK

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

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

RISL (SEQ ID NO: 26)

LLKK (SEQ ID NO: 27)

IPACTA (SEQ ID NO: 28)

FRAGGK (SEQ ID NO: 29)

HQYFA (SEQ ID NO: 30)

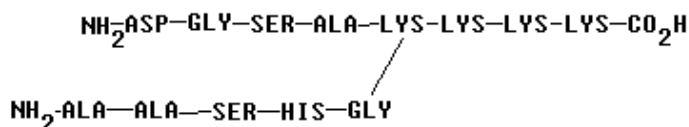
ATFGKKKA (SEQ ID NO: 31)

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

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

Example 7(b)-3: Branched amino acid sequence

Peptide of the following sequence:

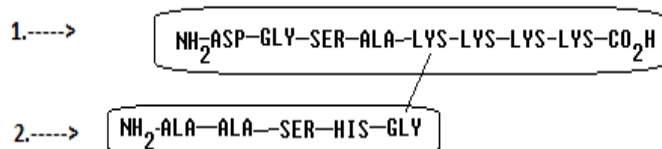


The linkage between the terminal Glycine residue in the lower sequence and the Lysine in the upper sequence is through an amide bond between the carboxy terminus of the Glycine and the amino terminal side chain of the Lysine.

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

YES

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



ST.26 paragraph 7(b) requires inclusion of peptides 1 and 2 in a sequence listing.

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

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

DGSAKKK (SEQ ID NO: 32)

AASHG (SEQ ID NO: 33)

Preferably the sequence DSAKXXX should include an annotation to indicate that the 5th lysine is a modified amino acid using the feature key "SITE" together with the qualifier "NOTE" describing that lysine links the peptide AASHG. Preferably the sequence AASHG should include an annotation to indicate that the 5th glycine is linked to DGSAKKK using the feature key "SITE" together with the qualifier "NOTE".

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

Paragraph 11(a) – Double-stranded nucleotide sequence – fully complementary

Example 11(a)-1: Double-stranded nucleotide sequence – same lengths

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

3' –CCGGTTAACGCTA–5'

5' –GGCCAATTGCGAT–3'

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

YES

Each enumerated nucleotide sequence has more than 10 specifically defined nucleotides. At least one strand must be included in the sequence listing, because the two strands of this double-stranded nucleotide sequence are fully complementary to each other.

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

While the sequence of only one strand must be included in the sequence listing, the sequences of both strands may be included, each with its own sequence identification number.

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

The double-stranded DNA sequence must be represented either as a single sequence or as two separate sequences. Each sequence included in the sequence listing must be represented in the 5' to 3' direction and assigned its own sequence identification number.

atcgcaattggcc (top strand) (SEQ ID NO: 34)

and/or

ggccaattgcat (bottom strand) (SEQ ID NO: 35)

Relevant ST.26 paragraphs: Paragraphs 7(a), 11(a), and 13

Paragraph 11(b) – Double-stranded nucleotide sequence - not fully complementary

Example 11(b)-1: Double-stranded nucleotide sequence – different lengths

A patent application contains the following drawing and caption:

5' -tagttcattgactaaggctccccattgactaaggcgactagcattgactaaggcaagc-3'



The human gene ABC1 promoter region (top strand) bound by a PNA probe (bottom strand). Where “n” in the PNA probe is a universal PNA base selected from the group consisting of 5-nitroindole and 3-nitroindole.

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

YES – the ABC1 promoter region (top strand)

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

YES – the PNA probe (bottom strand)

The bottom strand must also be included in the sequence listing, with its own sequence identification number, because the two strands are not fully complementary to each other. The individual residues that comprise a PNA or “peptide nucleic acid” are considered nucleotides according to ST.26 paragraph 3(g). Therefore, the bottom strand has more than 10 enumerated and “specifically defined” nucleotides and is required to be included in a sequence listing.

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

The top strand must be included in a sequence listing as:

tagttcattgactaaggctccccattgactaaggcgactagcattgactaaggcaagc (SEQ ID NO: 36)

The bottom strand is a peptide nucleic acid and therefore does not have a 3' and 5' end. According to paragraph 11, it must be included in a sequence listing “in the direction from left to right that mimics the 5'–end to 3'–end direction.” Therefore, it must be included in a sequence listing as:

cgcctnagtcaatggg (SEQ ID NO: 37)

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

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

Relevant ST.26 paragraphs: Paragraphs 3(g), 7(a), **11(b)**, 17, and 18

Example 11(b)-2: Double-stranded nucleotide sequence – no base-pairing segment

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

3'-CCGGTTAGCTTATACGCTAGGGCTA-5'

||||| |||||||||

5'-GGCCAATATGGCTTGCATCCCGAT-3'

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

YES

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

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

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

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

and

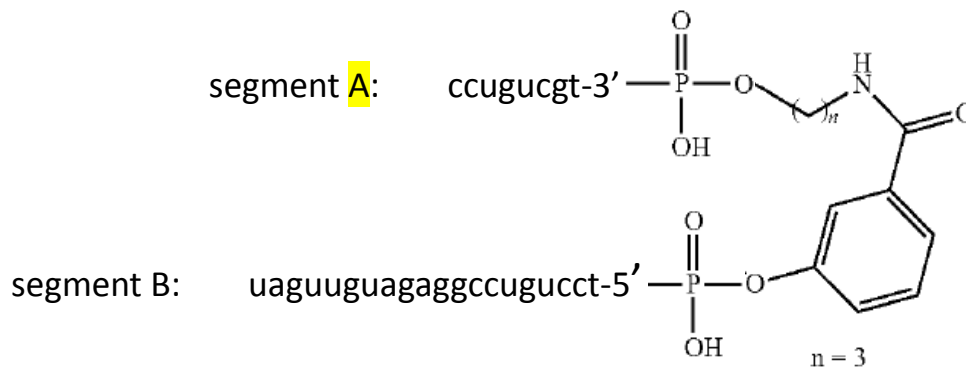
ggccaatatggcttgcgatcccgat (bottom strand) (SEQ ID NO: 39)

Relevant ST.26 paragraphs: Paragraphs 7(a), **11(b)**, and 13

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

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

A patent application describes the following compound:



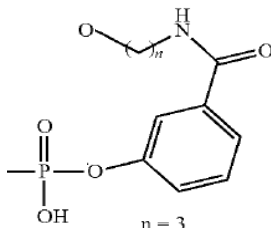
Wherein segment A and segment B are RNA sequences.

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

YES – segment B

NO – segment A

The enumerated sequence contains two segments of specifically defined nucleotides separated by the following "linker" structure:



The linker structure is not a nucleotide according to paragraph 3(g); therefore, each segment must be considered a separate sequence. Segment B contains more than 10 specifically defined nucleotides and ST.26 paragraph 7(a) requires inclusion in a sequence listing. Segment A contains only 8 specifically defined nucleotides and therefore is not required to be included in a sequence listing.

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

Segment A contains fewer than 10 specifically defined nucleotides, and therefore it must not be included in a sequence listing.

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

Segment B is an RNA molecule; therefore, the element "INSDSeq_moltype" must be "RNA." The symbol "u" must not be used to represent uracil in an RNA molecule in a sequence listing. According to paragraph 14, the symbol "t" will be construed as uracil in RNA. Accordingly, segment B must be included in the sequence listing as:

tcctgtccggagatgttgat (SEQ ID NO: 40)

Thymine in RNA is considered a modified nucleotide, i.e. modified uracil, and must be represented in the sequence as "t" and be further described in a feature table. Accordingly, the thymine in position 1 must be further described using the feature key "modified_base", the qualifier "mod_base" with "OTHER" as the qualifier value, and a qualifier "note" with "thymine" as the qualifier value.

The thymine, i.e. modified uracil, in position 1 should also be further described in a feature table using the feature key "misc_feature" and a qualifier "note" with the value e.g., "ccugucgt (Segment A) is attached at its 3'-end to a linker which is attached to the 5' oxygen of the thymidine. The linker is (4-(3-hydroxybenzamido)butyl) phosphinic acid."

Relevant ST.26 paragraphs: Paragraphs 3(g), 7(a), 8, 13, 14, 19, and 54

Paragraph 27 – The most restrictive ambiguity symbol should be used

Example 27-1: Shorthand formula for a nucleotide sequence

(GGGz)₂

Where z is any amino acid.

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

YES

The sequence is disclosed as a formula. (GGGz)₂ is simply a shorthand way of representing the sequence GGGzGGGz. Conventionally, a sequence is expanded first, and the definition of any variable, i.e. “z”, is determined thereafter.

The sequence uses the nonconventional symbol “z”. The definition of “z” must be determined from the explanation of the sequence in the disclosure, which defines this symbol as any amino acid (see Introduction to this document). The example does not provide any constraint on “z”, e.g., that it is the same in each occurrence.

Therefore, “z” is equivalent to the conventional symbol “X”, and the peptide in the example has eight enumerated amino acids, six of which are specifically defined glycine residues. ST.26 paragraph 7(b) requires inclusion of the sequence in a sequence listing as a single sequence with a single sequence identification number.

Note that the sequence is still encompassed by Paragraph 7(b) despite the fact that the enumerated and specifically defined residues are not contiguous.

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

The sequence uses the nonconventional symbol “z”, which according to the disclosure is any amino acid. The conventional symbol used to represent “any amino acid” is “X”. Therefore, the sequence must be represented as the single expanded sequence:

GGGXGGGX (SEQ ID NO: 41)

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

Relevant ST.26 paragraph(s): Paragraphs 3(c), 7(b) and 27

Example 27-2: Shorthand formula - less than four specifically defined amino acids

A peptide of the formula (Gly-Gly-Gly-z)_n

The disclosure further states, that z is any amino acid and

- (i) variable n is any length; or
- (ii) variable n is 2-100, preferably 3

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

NO

Consideration of both disclosed embodiments (i) and (ii) of the enumerated peptide of the formula reveals that “n” can be “any length”; therefore, the most encompassing embodiment of “n” is indeterminate. Since “n” is indeterminate, the peptide of the formula cannot be expanded to a definite length, and therefore, the unexpanded formula must be considered.

The enumerated peptide in the unexpanded formula (“n” = 1) provides three specifically defined amino acids, each of which is Gly, and the symbol “z”. Conventionally “Z” is the symbol for “glutamine or glutamic acid”; however, the example defines “z” as “any amino acid” (see Introduction to this document). Under ST.26, an amino acid that is not specifically defined is represented by “X”. Based on this analysis, the enumerated peptide, i.e. GGGX, does not

contain four specifically defined amino acids. Therefore, ST.26 paragraph 7(b) does not require inclusion, despite the fact that “n” is also defined as specific numerical values in some embodiments.

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

YES

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

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

A sequence containing 100 copies of GGGX is preferred (SEQ ID NO: 42). A further annotation should indicate that up to 98 copies of GGGX could be deleted. Inclusion of further specific embodiments that are a key part of the invention is strongly encouraged.

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

Relevant ST.26 paragraph(s): Paragraphs 3(c), 7(b), 26, and 27

Example 27-3: Shorthand formula - four or more specifically defined amino acids

A peptide of the formula (Gly-Gly-Gly-z)_n

Where z is any amino acid and variable n is 2-100, preferably 3.

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

YES

The enumerated peptide of the formula provides three specifically defined amino acids, each of which is Gly, and the symbol “z”. Conventionally, “Z” is the symbol for “glutamine or glutamic acid”; however, the description in this example defines “z” as “any amino acid” (see Introduction to this document). Under ST.26, an amino acid that is not specifically defined is represented by “X”. Based on this analysis, the enumerated repeat peptide does not contain four specifically defined amino acids. However, the description provides a specific numerical value for variable “n,” i.e., a lower limit of 2 and an upper limit of 100. Therefore, the example discloses a peptide having at least six specifically defined amino acids in the sequence GGGzGGGz, which is required by ST.26 to be included in a sequence listing.

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

Since “z” represents any amino acid, the conventional symbol used to represent the fourth and eighth amino acids is “X.”

ST.26 requires inclusion in a sequence listing of only the single sequence that has been enumerated by its residues. Therefore, at least one sequence containing any of 2, 3, or 100 copies of GGGX must be included in the sequence listing; however, the most encompassing sequence containing 100 copies of GGGX is preferred (SEQ ID NO: 42) (see Introduction to this document). In the latter case, a further annotation could indicate that up to 98 copies of GGGX could be deleted. Inclusion of two additional sequences containing 2 and 3 copies of GGGX, respectively (SEQ ID NO: 44-45), is strongly encouraged.

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

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

Relevant ST.26 paragraph(s): Paragraphs 3(c), 7(b), 26, and 27

Paragraph 28 – Amino acid sequences separated by internal terminator symbols

Example 28-1: Encoding nucleotide sequence and encoded amino acid sequence

A patent application describes the following sequences:

caattcaggg tggatgaat atg gcg ccc aat acg caa acc gcc tct ccc cgc
Met Ala Pro Asn Thr Gln Thr Ala Ser Pro Arg

gcg ttg gca gat tca tta atg cag ctg gca cga cag ggt tcc cga ctg
Ala Leu Ala Asp Ser Leu Met Gln Leu Ala Arg Gln Val Ser Arg Leu

Protein A

gaa agc ggg cag tga atg acc atg att acg gat tca ctg gcc gtc gtt
Glu Ser Gly Gln Met Thr Met Ile Thr Asp Ser Leu Ala Val Val

tta caa cgt cgt gac tgg gaa aac cct ggc ggt acc caa ctt aat cgc
Leu Gln Arg Arg Asp Trp Glu Asn Pro Gly Val Thr Gln Leu Asn Arg

Protein B

ctt gca gca cat tgg tgt caa aaa taa taataaccgg atgtactatt
Leu Ala Ala His Trp Cys Gln Lys

tatccctg atg ctg cgt cgt cag gtg aat gaa gtc gct taa gcaatcaatg
Met Leu Arg Arg Gln Val Asn Glu Val Ala

Protein C

tccgatgagg cgcgacgctt atccgaccaa catatcataa

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

YES

The application describes a nucleotide sequence, containing termination codons, which encodes three distinct amino acids sequences.

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

Regarding the encoded amino acid sequences, paragraph 28 requires that amino acid sequences separated by an internal terminator symbol such as a blank space, must be included as separate sequences. Since each of "Protein A", "Protein B", and "Protein C" contain four or more specifically defined amino acids, ST.26 paragraph 7(b) requires that each must be included in a sequence listing and must be assigned its own sequence identification number.

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

The nucleotide sequence must be included in a sequence listing as:

caattcaggggtggatgaatggcgcccaatcgcgcaaacgcctctccccgcgcttggccgattcattaatggaaagcgggcagtgatgaccatgattacggat
tcactggccgctgtttacaacgctgactgggaaaaccctggcgttacccaactaatcgccttgacgacacattggtgtcaaaaataataataaccggatgtacta

tttatccctgatgctgcgctgcaggtgaatgaagtcgctaagcaatcaatgctggatgcggcgacgcttatccgaccaacatatcataa. (SEQ ID NO: 46)

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

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

MAPNTQTASPRALADSLMQLARQVSRLESGQ (SEQ ID NO: 47)

MTMITDSLAVVLQRRDWENPGVTQLNRLAAHWCQK (SEQ ID NO: 48)

MLRRQVNEVA (SEQ ID NO: 49)

NOTE: See “Example 90-1 Amino acid sequence encoded by a coding sequence with introns” for an illustration of a translated amino acid sequence represented as a single sequence.

Relevant ST.26 paragraphs: Paragraphs 7, 26, 28, 57, 87-90

Paragraph 29 – Representation of an “other” amino acid

Example 29-1: Most restrictive ambiguity symbol for an “other” amino acid

A patent application describes the following sequence:

Ala-Hse-X₁-X₂-X₃-X₄-Tyr-Leu-Gly-Ser

Wherein, X₁= Ala or Gly,

X₂= Ala or Gly,

X₃= Ala or Gly,

X₄= Ala or Gly, and

Hse = Homoserine

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

YES

The enumerated peptide contains five specifically defined amino acids. The symbol “X” is used conventionally to represent two amino acids in the alternative (see Introduction to this document).

Because there are five specifically defined amino acids, i.e., Ala, Tyr, Leu, Gly and Ser, ST.26 paragraph 7(b) requires that the sequence must be included in a sequence listing.

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

Paragraph 29 requires any “other” amino acid must be represented by the symbol “X”. In the example, the sequence contains the amino acid Hse in position 2 which is not found in Annex I, Section 3, Table 3. Accordingly, Hse is an “other” amino acid and must be represented by the symbol “X”.

X₁-X₄ are variant positions, each of which can be A or G. The most restrictive ambiguity symbol for alternatives A or G is “X”. Therefore, the sequence may be represented as:

AXXXXXYLGS (SEQ ID NO: 50)

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

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

According to paragraph 27, because X₁-X₄ represent an alternative of only 2 amino acids, then further description is required. Paragraph 94 indicates that the feature key "VARIANT" should be used with the qualifier "NOTE" and qualifier value "A or G". According to ST.26 paragraph 34, since these positions are adjacent and have the same description, they may be jointly described using the syntax "3..6" as the location descriptor in the element INSDFeature_location.

Relevant ST.26 paragraphs: Paragraphs 3(a), 7(b), 25-27, **29**, 34, 66, 70, 71, and 94-95

Paragraph 30 – Annotation of a modified amino acid

Example 30-1 – Feature key "CARBOHYD"

A patent application describes a polypeptide with a specifically modified amino acid, containing a glycosylated side chain, characterized in that Cys corresponding to positions 4 and 15 of the polypeptide forms a disulfide bond, according to the following sequence:

Leu-Glu-Tyr-Cys-Leu-Lys-Arg-Trp-Asn(asialyloligosaccharide)-Glu-Thr-Ile-Ser-His-Cys-Ala-Trp

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

YES

The enumerated peptide provides 17 specifically defined amino acids. There are 16 natural amino acids, wherein the ninth (asparagine) is glycosylated. Therefore, the sequence must be included in a sequence listing as required by ST.26 paragraph (7)(b).

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

According to ST.26 paragraph 29, a modified amino acid should be represented in the sequence as the corresponding unmodified amino acid whenever possible.

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

LEYCLKRWNETISHCAW (SEQ ID NO: 51)

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

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

Relevant ST.26 paragraph(s): Paragraphs 3(a), 7(b), 26, 29, **30**, and Annex I, section 7, feature key 7.4

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

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

LL-100-KYMR

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

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

YES

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

The disclosed sequence uses a nonconventional symbol, i.e. “-100-.” The definition of “-100-” must be determined from the explanation of the sequence in the disclosure, which defines this symbol as 100 amino acids between leucine and lysine (see Introduction to this document). Therefore, “-100-” is a defined region of “X” residues. Since six of the 106 amino acids in the sequence are specifically defined, ST.26 paragraph 7(b) requires that the sequence must be included in a sequence listing.

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

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

Relevant ST.26 paragraph(s): Paragraphs 7(b), 26, 27, and 36

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

Lys-z₂-Lys-z_m-Lys-z₃-Lys-z_n-Lys-z₂-Lys

Where z is any amino acid, m=20, n=19-20, z₂ means that the pairs of Lysines are separated by any two amino acids, and z₃ means the pairs of Lysines are separated by any three amino acids.

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

YES

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

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

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

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

KXXKXXXXXXXXXXXXXXXXXXXXXXXXKXXKXXXXXXXXXXXXXXXXXXXXXXXXKXXK (SEQ ID NO: 53)

Wherein z_n is equal to 20 “X’s”, with a further description that the “X” variable corresponding to position 30 can be deleted.

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

KXXKXXXXXXXXXXXXXXXXXXXXXXXXKXXKXXXXXXXXXXXXXXXXXXXXXXXXKXXK (SEQ ID NO: 54)

Wherein z_n is equal to 19 “X’s”, with a further description that an “X” variable between position numbers 29 and 30 can be inserted.

Relevant ST.26 paragraph(s): Paragraphs 26, 27, and 36

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

K-z₂-K-z_m-K-z₃-K-z_n-K-z₂-K

Where z is any amino acid, where m=15-25, preferably 20-22, n=15-25, preferably 19-20, z₂ means that the pairs of Lysines are separated by any two amino acids, and z₃ means the pairs of Lysines are separated by any three amino acids.

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

YES

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

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

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

KXXKXXXXXXXXXXXXXXXXXXXXXXXXXXXXKXXKXXXXXXXXXXXXXXXXXXXXXXXXXXXXKXXK (SEQ ID NO: 55)

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

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

Alternatively, the sequence may be represented as:

KXXKXXXXXXXXXXXXXXXXKXXKXXXXXXXXXXXXXXXXKXXK (SEQ ID NO: 56)

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

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

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

Relevant ST.26 paragraph(s): Paragraphs 27 and 36

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

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

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

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

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

NO

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

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

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

NO – not the entire sequence

NO – not any portion of the sequence

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

ST.26 paragraph 8 does not permit inclusion of either portion adjacent to the undefined gap, since each portion contains only two specifically defined amino acids.

Relevant ST.26 paragraphs: Paragraphs 7(b), 8, 26, and 37

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

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

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

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

NO – not the entire sequence

YES – a portion of the sequence

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

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

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

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

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

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

GGAGXX (SEQ ID NO: 57)

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

Relevant ST.26 paragraph(s): Paragraphs 7(b), 8, 26, and 37

Paragraph 87 – "CDS" Feature key

Example 87-1: Encoding nucleotide sequence and encoded amino acid sequence

A patent application describes the following nucleotide sequence and its translation:

atg acc gga aat aaa cct gaa acc gat gtt tac gaa att tta tga

Met Thr Gly Asn Lys Pro Glu Thr Asp Val Tyr Glu Ile Leu STOP

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

YES – the nucleotide sequence. The enumerated nucleotide sequence has more than ten specifically defined nucleotides.

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

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

The nucleotide sequence must be presented as:

atgaccggaataaacctgaaaccgatgtttacgaaattttatga (SEQ ID NO: 58)

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

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

MTGNKPETDVYEIL (SEQ ID NO: 59)

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

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

Relevant ST.26 paragraphs: Paragraphs 7(a), 7(b), 26, 28, 87, 88, and 90

Paragraph 90 – Amino acid sequence encoded by a coding sequence

Example 90-1: Amino acid sequence encoded by a coding sequence with introns

A patent application contains the following figure disclosing a coding sequence and its translation:

```
atg aag act ttc gca gcc ttg ctt tcc gct gtc act ctc gcg ctc tcg
Met Lys Thr Phe Ala Ala Leu Leu Ser Ala Val Thr Leu Ala Leu Ser

gtg cgc gcc cag gcg gct gtc tgg agt caa t gtaagtgccg ctgcttttca
Val Arg Ala Gln Ala Ala Val Trp Ser Gln

ttgatacgag actctacgcc gagctgacgt gctaccgtat ag gt ggc ggt aca
Cys Gly Gly Thr

ccg ggt tgg acg gcc gag acc act tgc gtt gct ggt tcg gtt tgt acc
Pro Gly Trp Thr Gly Glu Thr Thr Cys Val Ala Gly Ser Val Cys Thr

tcc ttg agc tca gtgagcgact ttcaatccgt cgtcattgct cctcatgtat
Ser Leu Ser Ser

tgacgattgg ccttcatag tca tac tct caa tgc gtt ccg gcc tcc gca acg
Ser Tyr Ser Gln Cys Val Pro Gly Ser Ala Thr

tcc agc gct ccg gcg gcc ccc tca gcg aca act tca gcc ccc gca cct
Ser Ser Ala Pro Ala Ala Pro Ser Ala Thr Thr Ser Gly Pro Ala Pro

acg gac gga acg tgc tcg gcc agc ggg gca tgg ccg cca ttg acc tga
Thr Asp Gly Thr Cys Ser Ala Ser Gly Ala Trp Pro Pro Leu Thr Ter
```

Figure 1 – nucleotides shown in bold-face are intron regions.

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

YES

The application discloses a nucleotide sequence and its amino acid translation. The enumerated nucleotide sequence contains more than 10 specifically defined nucleotides and must be included in a sequence listing as a single sequence.

The nucleotide sequence contains coding sequence (exons) separated by noncoding sequence (introns). The figure depicts the translation of the nucleotide sequence as three non-contiguous amino acid sequences. According to the figure caption, the bolded regions of nucleotides are intron sequences that will be spliced out of an RNA transcript before translation into a protein. Accordingly, the three amino acid sequences are actually a single, contiguous, enumerated sequence, which contains more than four specifically defined amino acids and must be included in a sequence listing as a single sequence.

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

The nucleotide sequence must be included in a sequence listing as:

```
atgaagactttcgcagccttgccttcgctgtcactctcgcgctcgcggtgcgcgcccagggcggtgctctggagtcagtgaagtgccgctgctttcattgatacgaga
ctctacgccgagctgacgtgctaccgtataggtggcggtacaccgggtggacggcgagaccactgctgtgctggttcggtttgtacctcttgagctcagtgag
cgacttcaatccgctcattgctcctcatgtattgacgattggcctcatagtcatactctcaatgcgttccgggctccgcaacgtccagcctccggcggccccctc
agcgacaactcaggccccgcactacggacggaacgtgctcggccagcggggcatggccgcatgacctga (SEQ ID NO: 74)
```

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

The amino acid sequence must be included as a single sequence:

```
MKTFAALLSAVTLALSVRAQAAVWSQCGGTPGWTGETTCVAGSVCTSLSSSYSQCVPGSATSSAPAAPSATTSG
PAPTDGTCSASGAWPPLT (SEQ ID NO: 75)
```

Relevant ST.26 paragraphs: Paragraphs 7, 26, 28, 57, 87-90

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

Example 91-1: Representation of enumerated variants

The description includes the following sequence alignment.

```
D. melanogaster      ACATTGAATCTCATACCACTTT
D. virilis           ...-..G...C...-..G....
D. simulans         GT..G.CG..GT..SGT.G...
```

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

YES

It is common in the art to include “dots” in a sequence alignment to indicate “this position is the same as the position above it.” Therefore, the “dots” in species 2 and 3 are considered enumerated and specifically defined nucleotides, as they are simply a short-hand way of indicating that a given position is the same nucleotide as in species 1. In addition, sequence alignments frequently display the symbol “-” to indicate the absence of a residue in order to maximize the alignment.

Accordingly, the nucleotide sequences of species 1 and 3 contain twenty-two enumerated and specifically defined nucleotides, whereas the nucleotide sequences of species 2 contains nineteen. Thus, each sequence is required by ST.26 paragraph 7(a) to be included in a sequence listing with separate sequence identification numbers.

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

Drosophila melanogaster sequence must be included in a sequence listing as:

```
acattgaatctataccacttt (SEQ ID NO: 60)
```

Drosophila virilis sequence must be included in a sequence listing as:

```
acatggatcccacgacttt (SEQ ID NO: 61)
```

Drosophila simulans sequence must be included in a sequence listing as:

```
gtagggcgtcgtatsgtagttt (SEQ ID NO: 62)
```

Relevant ST.26 paragraphs: Paragraphs 7(a), 13, and 91

Example 91-2: Representation of enumerated variants

The description includes the following table of a peptide and functional variants thereof. A blank space in the table below indicates that an amino acid in the variant is the same as the corresponding amino acid in the "Sequence" and a "-" indicates deletion of the corresponding amino acid in the "Sequence".

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

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

YES

As indicated, a blank space in this table indicates that an amino acid in the variant is the same as the corresponding amino acid in the "Sequence". Therefore, the amino acids of the variant sequences are enumerated and specifically defined.

Since the four variant sequences each contain more than four enumerated and specifically defined amino acids, each sequence is required by ST.26 paragraph 7(a) to be included in a sequence listing with separate sequence identification numbers.

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

AVLTYLRGE (SEQ ID NO: 76)

AVLTYLRGA (SEQ ID NO: 77)

AVPTYPRGE (SEQ ID NO: 78)

AVAIGYRGE (SEQ ID NO: 79)

AVLTYLGE (SEQ ID NO: 80)

Relevant ST.26 paragraphs: Paragraphs 7(b), 26, and 91

Example 91-3: Representation of a consensus sequence

A patent application includes Figure 1 with the following multiple sequence alignment.

```

Consensus      LEGnEQFINAAkIIRHPkYnrkTlNNDIMLIK
Homo sapiens   LEGNEQFINAAKIIRHPQYDRKTLNNDIMLIK
Pongo abelii   LEGNEQFINAAKIIRHPQYDRKTVNNDIMLIK
Papio anubis   LEGTEQFINAAKIIRHPDYDRKTLNNDILLIK
Rhinopithecus roxellana LEGTEQFINAAKIIRHPNYNRITLDNDILLIK
Pan paniscus   LEGNEQFINAAKIIRHPKYNRITLNDIMLIK
Rhinopithecus bieti LEGNEQFINATKIIRHPKYNGNTLNNDIMLIK
Rhinopithecus roxellana LEGNEQFINATQIIRHPKYNGNTLNNDIMLIK
  
```

The consensus sequence includes upper case letters to represent conserved amino acid residues, while the lower case letters “n”, “a”, “k”, “r”, “l” and “m” represent the predominant amino acid residues among the aligned sequences.

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

YES

The lower case letters in the consensus sequence each represent a single amino acid residue. Consequently, the consensus sequence, as well as each of the remaining seven sequences in Figure 1, includes at least four specifically defined amino acids. ST.26 paragraph 7(b) requires inclusion of all eight sequences in the sequence listing.

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

The lower case letters in the consensus sequence are being used as ambiguity symbols to represent the predominant amino acid among the possible variants for a specific position. Therefore, the lower case letters “n”, “a”, “k”, “r”, “l” and “m” are conventional symbols used in a nonconventional manner and the consensus sequence must be represented using an ambiguity symbol in place of each of the lower case letters.

The most restrictive ambiguity symbol should be used. For most positions in the consensus sequence, “X” is the most restrictive ambiguity symbol; however, the most restrictive ambiguity symbol for “D” or “N” in positions 20 and 25 is “B”. The consensus sequence should be included in the sequence listing as:

LEGXEQFINAXXIIRHPXYBXXTXBNIDXLIK (SEQ ID NO: 81)

According to paragraph 27, the symbol “X” will be construed as any one of “A”, “R”, “N”, “D”, “C”, “Q”, “E”, “G”, “H”, “I”, “L”, “K”, “M”, “F”, “P”, “O”, “S”, “U”, “T”, “W”, “Y”, or “V”, except where it is used with a further description in the feature table. Therefore, each “X” in the consensus sequence must be further described in a feature table using the feature key “VARIANT” and the qualifier “NOTE” to indicate the possible variants for each position.

The remaining seven sequences must be included in the sequence listing as:

LEGNEQFINAAKIIRHPQYDRKTLNNDIMLIK (SEQ ID NO: 82)

LEGNEQFINAAKIIRHPQYDRKTVNNDIMLIK (SEQ ID NO: 83)

LEGTEQFINAAKIIRHPDYDRKTLNNDILLIK (SEQ ID NO: 84)

LEGTEQFINAAKIIRHPNYNRITLDNDILLIK (SEQ ID NO: 85)

LEGNEQFINAAKIIRHPKYNRITLNDIMLIK (SEQ ID NO: 86)

LEGNEQFINATKIIRHPKYNGNTLNNDIMLIK (SEQ ID NO: 87)

LEGNEQFINATQIIRHPKYNGNTLNNDIMLIK (SEQ ID NO: 88)

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

Relevant ST.26 paragraphs: Paragraphs 7(b), 26, 27, 91, and 95

Paragraph 92 – Variant sequence disclosed as a single sequence with enumerated alternative residues

Example 92-1: Representation of single sequence with enumerated alternative amino acids

A patent application claims a peptide of the sequence:

(i) Gly-Gly-Gly-[Leu or Ile]-Ala-Thr-[Ser or Thr]

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

YES

The sequence provides four specifically defined amino acids and ST.26 paragraph 7(b) requires inclusion of the sequence in a sequence listing.

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

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

GGGJATX (SEQ ID NO: 63)

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

Alternatively, the sequence may be represented, for example, as:

GGGLATS (SEQ ID NO: 64)

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

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

Relevant ST.26 paragraph(s): Paragraphs 7(b), 8, 26, 27, **92**, and 95

Paragraph 93(a) – A variant sequence disclosed only by reference to a primary sequence with multiple independent variations

Example 93(a)-1: Representation of a variant sequence by annotation of the primary sequence

An application contains the following disclosure:

“Peptide fragment 1 is Gly-Leu-Pro-Xaa-Arg-Ile-Cys wherein Xaa can be any amino acid....

In another embodiment, peptide fragment 1 is Gly-Leu-Pro-Xaa-Arg-Ile-Cys wherein Xaa can be Val, Thr, or Asp....

In another embodiment, peptide fragment 1 is Gly-Leu-Pro-Xaa-Arg-Ile-Cys wherein Xaa can be Val.”

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

YES

“Peptide fragment 1” in each of the three disclosed embodiments provides at least six specifically defined amino acids; therefore, the sequence must be included in a sequence listing as required by ST.26 paragraph 7(b).

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

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

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

GLPXRIC (SEQ ID NO: 65)

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

For the above example, it is strongly encouraged that the following additional three sequences are included in the sequence listing, each with their own sequence identification number:

GLPVRI (SEQ ID NO: 66)

GLPTRIC (SEQ ID NO: 67)

GLPDRI (SEQ ID NO: 68)

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

Relevant ST.26 paragraph(s): Paragraphs 7(b), 26, 27, and 93(a)

Paragraph 93(b) – A variant sequence disclosed only by reference to a primary sequence with multiple interdependent variations

Example 93(b)-1: Representation of individual variant sequences with multiple interdependent variations

A patent application describes the following consensus sequence:

cgaatg n_1 cccactacgaatg n_2 cacgaatg n_3 cccaca

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

Several variant sequences are disclosed as follows:

if n_1 is a, then n_2 and n_3 are t, g, or c;

if n_1 is t, then n_2 and n_3 are a, g, or c;

if n_1 is g, then n_2 and n_3 are t, a, or c;

if n_1 is c, then n_2 and n_3 are t, g, or a.

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

YES

The sequence has more than ten enumerated and “specifically defined” nucleotides and is required by ST.26 paragraph 7(a) to be included in a sequence listing.

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

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

cgaatg n cccactacgaatg n cacgaatg n cccaca (SEQ ID NO: 69)

The enumerated sequence contains variations at three distinct locations and the occurrence of the variations is interdependent. Inclusion of additional sequences which represent additional embodiments that are a key part of the invention is **strongly** encouraged, as discussed in the introduction to this document. Therefore, according to ST.26 paragraph 93(b), the additional embodiments should be included in a sequence listing as four separate sequences, each with its own sequence identification number:

cgaatg a cccactacgaatg b cacgaatg b cccaca (SEQ ID NO: 70)

cgaatg t cccactacgaatg v cacgaatg v cccaca (SEQ ID NO: 71)

cgaatg g cccactacgaatg h cacgaatg h cccaca (SEQ ID NO: 72)

cgaatg c cccactacgaatg d cacgaatg d cccaca (SEQ ID NO: 73)

(Note that b = t, g, or c; v = a, g, or c; h = t, a, or c; and d = t, g, or a; see Annex I, Section 1, Table 1)

According to ST.26 paragraph 15, the most restrictive symbol must be used to represent variable positions. Consequently, n_2 and n_3 must not be represented by “n” in the sequence.

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

Relevant ST.26 paragraphs: Paragraphs 7(a), 15, and 93(b)

Appendix

GUIDANCE DOCUMENT SEQUENCES IN XML

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ST26SequenceListing PUBLIC "-//WIPO//DTD Sequence Listing 1.0//EN"
"resources/ST26SequenceListing_V1_0.dtd">
<ST26SequenceListing dtdVersion="1" fileName="Guidance_Document_Sequences_XML" softwareName="prototype"
softwareVersion="1.0" productionDate="2017-01-02">
  <ApplicantFileReference>ABCD#1234</ApplicantFileReference>
  <ApplicantName languageCode="de">Johannes Jäger</ApplicantName>
  <ApplicantNameLatin>Johannes Jaeger</ApplicantNameLatin>
  <InventionTitle languageCode="de">Pharmakologische Wirkstoffe für das Nervensystem</InventionTitle>
  <SequenceTotalQuantity>88</SequenceTotalQuantity>
  <SequenceData sequenceIDNumber="1">
    <INSDSeq>
      <INSDSeq_length>7</INSDSeq_length>
      <INSDSeq_moltype>AA</INSDSeq_moltype>
      <INSDSeq_division>PAT</INSDSeq_division>
      <INSDSeq_feature-table>
        <INSDFeature>
          <INSDFeature_key>SOURCE</INSDFeature_key>
          <INSDFeature_location>1..7</INSDFeature_location>
          <INSDFeature_qual>
            <INSDQualifier>
              <INSDQualifier_name>ORGANISM</INSDQualifier_name>
              <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
            </INSDQualifier>
            <INSDQualifier>
              <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
              <INSDQualifier_value>protein</INSDQualifier_value>
            </INSDQualifier>
          </INSDFeature_qual>
        </INSDFeature>
        <INSDFeature>
          <INSDFeature_key>REGION</INSDFeature_key>
          <INSDFeature_location>1..7</INSDFeature_location>
          <INSDFeature_qual>
            <INSDQualifier>
              <INSDQualifier_name>NOTE</INSDQualifier_name>
              <INSDQualifier_value>circular peptide</INSDQualifier_value>
            </INSDQualifier>
          </INSDFeature_qual>
        </INSDFeature>
        <INSDFeature>
          <INSDFeature_key>SITE</INSDFeature_key>
          <INSDFeature_location>1</INSDFeature_location>
          <INSDFeature_qual>
            <INSDQualifier>
              <INSDQualifier_name>NOTE</INSDQualifier_name>
              <INSDQualifier_value>D-Alanine</INSDQualifier_value>
            </INSDQualifier>
          </INSDFeature_qual>
        </INSDFeature>
        <INSDFeature>
          <INSDFeature_key>SITE</INSDFeature_key>
          <INSDFeature_location>2</INSDFeature_location>
          <INSDFeature_qual>
            <INSDQualifier>
              <INSDQualifier_name>NOTE</INSDQualifier_name>
              <INSDQualifier_value>D-Glutamic acid</INSDQualifier_value>
            </INSDQualifier>
          </INSDFeature_qual>
        </INSDFeature>
        <INSDFeature>
          <INSDFeature_key>SITE</INSDFeature_key>
          <INSDFeature_location>4</INSDFeature_location>
          <INSDFeature_qual>
            <INSDQualifier>
              <INSDQualifier_name>NOTE</INSDQualifier_name>
              <INSDQualifier_value>Nle</INSDQualifier_value>
            </INSDQualifier>
          </INSDFeature_qual>
        </INSDFeature>
      </INSDSeq_feature-table>
    </INSDSeq>
  </SequenceData>
</ST26SequenceListing>
</?xml>
```

```
        </INSDFeature_qual>
    </INSDFeature>
    <INSDFeature>
        <INSDFeature_key>SITE</INSDFeature_key>
        <INSDFeature_location>6</INSDFeature_location>
        <INSDFeature_qual>
            <INSDQualifier>
                <INSDQualifier_name>NOTE</INSDQualifier_name>
                <INSDQualifier_value>D-Methionine</INSDQualifier_value>
            </INSDQualifier>
        </INSDFeature_qual>
    </INSDFeature>
    <INSDFeature>
        <INSDFeature_key>SITE</INSDFeature_key>
        <INSDFeature_location>7</INSDFeature_location>
        <INSDFeature_qual>
            <INSDQualifier>
                <INSDQualifier_name>NOTE</INSDQualifier_name>
                <INSDQualifier_value>D-Norleucine</INSDQualifier_value>
            </INSDQualifier>
        </INSDFeature_qual>
    </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>AEKXGMX</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="2">
    <INSDSeq>
        <INSDSeq_length>6</INSDSeq_length>
        <INSDSeq_moltype>AA</INSDSeq_moltype>
        <INSDSeq_division>PAT</INSDSeq_division>
        <INSDSeq_feature-table>
            <INSDFeature>
                <INSDFeature_key>SOURCE</INSDFeature_key>
                <INSDFeature_location>1..6</INSDFeature_location>
                <INSDFeature_qual>
                    <INSDQualifier>
                        <INSDQualifier_name>ORGANISM</INSDQualifier_name>
                        <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
                    </INSDQualifier>
                    <INSDQualifier>
                        <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
                        <INSDQualifier_value>protein</INSDQualifier_value>
                    </INSDQualifier>
                </INSDFeature_qual>
            </INSDFeature>
            <INSDFeature>
                <INSDFeature_key>SITE</INSDFeature_key>
                <INSDFeature_location>1</INSDFeature_location>
                <INSDFeature_qual>
                    <INSDQualifier>
                        <INSDQualifier_name>NOTE</INSDQualifier_name>
                        <INSDQualifier_value>the N-terminus is acetylated</INSDQualifier_value>
                    </INSDQualifier>
                </INSDFeature_qual>
            </INSDFeature>
            <INSDFeature>
                <INSDFeature_key>SITE</INSDFeature_key>
                <INSDFeature_location>4</INSDFeature_location>
                <INSDFeature_qual>
                    <INSDQualifier>
                        <INSDQualifier_name>NOTE</INSDQualifier_name>
                        <INSDQualifier_value>6-amino-7-(1H-indol-3-yl)-5-oxoheptanoic
acid</INSDQualifier_value>
                    </INSDQualifier>
                </INSDFeature_qual>
            </INSDFeature>
            <INSDFeature>
                <INSDFeature_key>SITE</INSDFeature_key>
                <INSDFeature_location>6</INSDFeature_location>
                <INSDFeature_qual>
                    <INSDQualifier>
                        <INSDQualifier_name>NOTE</INSDQualifier_name>
                        <INSDQualifier_value>the C-terminus is methylated</INSDQualifier_value>
                    </INSDQualifier>
                </INSDFeature_qual>
            </INSDFeature>
        </INSDSeq_feature-table>
    </INSDSeq>
</SequenceData>
```

```

        </INSDQualifier>
      </INSDFeature_qual>
    </INSDFeature>
  </INSDSeq_feature-table>
  <INSDSeq_sequence>VAFXGK</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="3">
  <INSDSeq>
    <INSDSeq_length>4</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..4</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>SITE</INSDFeature_key>
        <INSDFeature_location>1</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>the N-terminus is acetylated</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>SITE</INSDFeature_key>
        <INSDFeature_location>4</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>C-terminus linked via a glutaraldehyde bridge to
dipeptide GK</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>VAFW</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="4">
  <INSDSeq>
    <INSDSeq_length>5</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..5</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>REGION</INSDFeature_key>

```

```

    <INSDFeature_location>&gt;5</INSDFeature_location>
    <INSDFeature_qual>
      <INSDQualifier>
        <INSDQualifier_name>NOTE</INSDQualifier_name>
        <INSDQualifier_value>The entire sequence of amino acids 1-5 can be repeated
one or more times</INSDQualifier_value>
      </INSDQualifier>
    </INSDFeature_qual>
  </INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>GGGGX</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="5">
  <INSDSeq>
    <INSDSeq_length>12</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..12</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>12</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>cytosine attached to a C3 spacer, which is joined to
another nucleic acid</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>atgcatgcatgc</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="6">
  <INSDSeq>
    <INSDSeq_length>12</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..12</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</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>

```

```

    <INSDQualifier_value>cytosine attached to a C3 spacer, which is joined to
another nucleic acid</INSDQualifier_value>
  </INSDQualifier>
</INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>cggcatgcatgc</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="7">
  <INSDSeq>
    <INSDSeq_length>25</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..25</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>organism</INSDQualifier_name>
            <INSDQualifier_value>Mus musculus</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>atgcatgcatgcvcgcatgcatgc</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="8">
  <INSDSeq>
    <INSDSeq_length>12</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..12</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>12</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>cytosine is linked to a C3 spacer, which is linked to
5&apos;-end of another nucleic acid</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>atgcatgcatgc</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="9">
  <INSDSeq>
    <INSDSeq_length>12</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>

```

```

<INSDFeature>
  <INSDFeature_key>source</INSDFeature_key>
  <INSDFeature_location>1..12</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</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>note</INSDQualifier_name>
      <INSDQualifier_value>cytosine is linked to a C3 spacer, which is linked to
3&apos;-end of another nucleic acid</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>cggcatgcatgc</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="10">
  <INSDSeq>
    <INSDSeq_length>18</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..18</INSDFeature_location>
        <INSDFeature_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>modified_base</INSDFeature_key>
        <INSDFeature_location>11</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>mod_base</INSDQualifier_name>
            <INSDQualifier_value>OTHER</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>abasic site</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>gagcattgacntaaggct</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="11">
  <INSDSeq>
    <INSDSeq_length>30</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>

```



```

<INSDFeature_key>source</INSDFeature_key>
<INSDFeature_location>1..30</INSDFeature_location>
<INSDFeature_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>
  <INSDQualifier>
    <INSDQualifier_name>note</INSDQualifier_name>
    <INSDQualifier_value>GNA sequence</INSDQualifier_value>
  </INSDQualifier>
</INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>modified_base</INSDFeature_key>
  <INSDFeature_location>1..30</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>mod_base</INSDQualifier_name>
      <INSDQualifier_value>OTHER</INSDQualifier_value>
    </INSDQualifier>
    <INSDQualifier>
      <INSDQualifier_name>note</INSDQualifier_name>
      <INSDQualifier_value> 2,3-dihydroxypropyl nucleosides (glycol nucleic
acids)</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>tagttcattgactaaggctccccattgact</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="12">
  <INSDSeq>
    <INSDSeq_length>10</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..10</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>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>aatgccggag</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="13">
  <INSDSeq>
    <INSDSeq_length>10</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..10</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>organism</INSDQualifier_name>
            <INSDQualifier_value>synthetic construct</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
  </INSDSeq>
</SequenceData>

```

```

        <INSDQualifier>
          <INSDQualifier_name>mol_type</INSDQualifier_name>
          <INSDQualifier_value>other DNA</INSDQualifier_value>
        </INSDQualifier>
      </INSDFeature_qual>
    </INSDFeature>
  </INSDSeq_feature-table>
  <INSDSeq_sequence>aatgccggak</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="14">
  <INSDSeq>
    <INSDSeq_length>10</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..10</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>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>aatgccggac</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="15">
  <INSDSeq>
    <INSDSeq_length>10</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..10</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>aatgttgac</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="16">
  <INSDSeq>
    <INSDSeq_length>15</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..15</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>
</INSDSeq_feature-table>
<INSDSeq_sequence>nngkngkngkagvcr</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="17">
  <INSDSeq>
    <INSDSeq_length>7</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..7</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>XYEKGJL</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="18">
  <INSDSeq>
    <INSDSeq_length>30</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..30</INSDFeature_location>
        <INSDFeature_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</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>This sequence is one of four branches of a branched
polynucleotide</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>cacacaaaaaaaaaaaaaaaaaaaaaaaa</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="19">
  <INSDSeq>
    <INSDSeq_length>28</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>

```

```

    <INSDFeature_key>source</INSDFeature_key>
    <INSDFeature_location>1..28</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</INSDFeature_location>
    <INSDFeature_qual>
      <INSDQualifier>
        <INSDQualifier_name>note</INSDQualifier_name>
        <INSDQualifier_value>This sequence is one of four branches of a branched
polynucleotide.</INSDQualifier_value>
      </INSDQualifier>
    </INSDFeature_qual>
  </INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>cacataggcacatctccttagtgcaggaaga</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="20">
  <INSDSeq>
    <INSDSeq_length>75</INSDSeq_length>
    <INSDSeq_moltype>RNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..75</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>tRNA</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>modified_base</INSDFeature_key>
        <INSDFeature_location>39</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>mod_base</INSDQualifier_name>
            <INSDQualifier_value>p</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>modified_base</INSDFeature_key>
        <INSDFeature_location>54</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>mod_base</INSDQualifier_name>
            <INSDQualifier_value>p</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
  </INSDSeq>
</SequenceData>
<INSDSeq_sequence>gcggatttagctcagctgggagagcgcagactgaatanctggagtctgtgtncgatccacagaattcgacca</INSDSeq_
sequence>
</INSDSeq>
</SequenceData>

```

```
<SequenceData sequenceIDNumber="21">
  <INSDSeq>
    <INSDSeq_length>44</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..44</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>gatcattttttatattttttatattttttatattttttatgtac</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="22">
  <INSDSeq>
    <INSDSeq_length>44</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..44</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>repeat_region</INSDFeature_key>
        <INSDFeature_location>5..40</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>rpt_type</INSDQualifier_name>
            <INSDQualifier_value>tandem</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>rpt_unit_seq</INSDQualifier_name>
            <INSDQualifier_value>atcgcact</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>attcatcgcactatcgcactatcgcactatcgcactgtac</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="23">
  <INSDSeq>
    <INSDSeq_length>44</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..44</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>repeat_region</INSDFeature_key>
  <INSDFeature_location>5..40</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>rpt_type</INSDQualifier_name>
      <INSDQualifier_value>tandem</INSDQualifier_value>
    </INSDQualifier>
    <INSDQualifier>
      <INSDQualifier_name>rpt_unit_seq</INSDQualifier_name>
      <INSDQualifier_value>atacgcact</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>gatcatacgcactatacgcactatacgcactatacgcactgtac</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="24">
  <INSDSeq>
    <INSDSeq_length>12</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..12</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_difference</INSDFeature_key>
        <INSDFeature_location>5..8</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>A nucleotide in any of positions 5-8 may be replaced
with inosine or pseudouridine.</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>modified_base</INSDFeature_key>
        <INSDFeature_location>5..8</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>i or p</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>gatcaaaagtac</INSDSeq_sequence>
  </INSDSeq>

```



```
        <INSDQualifier_value>synthetic construct</INSDQualifier_value>
      </INSDQualifier>
    <INSDQualifier>
      <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
      <INSDQualifier_value>protein</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>1..4</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>This sequence is one part of a branched
peptide</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>LLKK</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="28">
  <INSDSeq>
    <INSDSeq_length>6</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..6</INSDFeature_location>
        <INSDFeature_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>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>SITE</INSDFeature_key>
        <INSDFeature_location>1..6</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>This sequence is one part of a branched
peptide</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>IPACTA</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="29">
  <INSDSeq>
    <INSDSeq_length>6</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..6</INSDFeature_location>
        <INSDFeature_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>
    </INSDFeature_qual>
  </INSDFeature>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>1..6</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>This sequence is one part of a branched
peptide</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>FRAGGK</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="30">
  <INSDSeq>
    <INSDSeq_length>5</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..5</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>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>SITE</INSDFeature_key>
        <INSDFeature_location>1..5</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>This sequence is one part of a branched
peptide</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>HQYFA</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="31">
  <INSDSeq>
    <INSDSeq_length>8</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..8</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>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
  </INSDSeq>
</SequenceData>

```

```

    <INSDFeature>
      <INSDFeature_key>SITE</INSDFeature_key>
      <INSDFeature_location>1..8</INSDFeature_location>
      <INSDFeature_qual>
        <INSDQualifier>
          <INSDQualifier_name>NOTE</INSDQualifier_name>
          <INSDQualifier_value>This sequence is one part of a branched
peptide</INSDQualifier_value>
        </INSDQualifier>
      </INSDFeature_qual>
    </INSDFeature>
  </INSDSeq_feature-table>
  <INSDSeq_sequence>ATFGKKA</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="32">
  <INSDSeq>
    <INSDSeq_length>7</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..7</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>SITE</INSDFeature_key>
        <INSDFeature_location>5</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>Bound to Gly at position 5 of
AASHG</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>DGSAKKK</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="33">
  <INSDSeq>
    <INSDSeq_length>5</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..5</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>SITE</INSDFeature_key>
        <INSDFeature_location>5</INSDFeature_location>
        <INSDFeature_qual>

```

```
                <INSDQualifier>
                  <INSDQualifier_name>NOTE</INSDQualifier_name>
                  <INSDQualifier_value>Bound to Lys at position 5 of
DGSAKKK</INSDQualifier_value>
                </INSDQualifier>
              </INSDFeature_qual>
            </INSDFeature>
          </INSDSeq_feature-table>
        <INSDSeq_sequence>AASHG</INSDSeq_sequence>
      </INSDSeq>
    </SequenceData>
  <SequenceData sequenceIDNumber="34">
    <INSDSeq>
      <INSDSeq_length>13</INSDSeq_length>
      <INSDSeq_moltype>DNA</INSDSeq_moltype>
      <INSDSeq_division>PAT</INSDSeq_division>
      <INSDSeq_feature-table>
        <INSDFeature>
          <INSDFeature_key>source</INSDFeature_key>
          <INSDFeature_location>1..13</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>atcgcaattggcc</INSDSeq_sequence>
    </INSDSeq>
  </SequenceData>
  <SequenceData sequenceIDNumber="35">
    <INSDSeq>
      <INSDSeq_length>13</INSDSeq_length>
      <INSDSeq_moltype>DNA</INSDSeq_moltype>
      <INSDSeq_division>PAT</INSDSeq_division>
      <INSDSeq_feature-table>
        <INSDFeature>
          <INSDFeature_key>source</INSDFeature_key>
          <INSDFeature_location>1..13</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>ggccaattgcat</INSDSeq_sequence>
    </INSDSeq>
  </SequenceData>
  <SequenceData sequenceIDNumber="36">
    <INSDSeq>
      <INSDSeq_length>58</INSDSeq_length>
      <INSDSeq_moltype>DNA</INSDSeq_moltype>
      <INSDSeq_division>PAT</INSDSeq_division>
      <INSDSeq_feature-table>
        <INSDFeature>
          <INSDFeature_key>source</INSDFeature_key>
          <INSDFeature_location>1..58</INSDFeature_location>
          <INSDFeature_qual>
            <INSDQualifier>
              <INSDQualifier_name>organism</INSDQualifier_name>
              <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
            </INSDQualifier>
            <INSDQualifier>
            </INSDQualifier>
          </INSDFeature_qual>
        </INSDFeature>
      </INSDSeq_feature-table>
    </INSDSeq>
  </SequenceData>
```

```

        <INSDQualifier_name>mol_type</INSDQualifier_name>
        <INSDQualifier_value>genomic DNA</INSDQualifier_value>
      </INSDQualifier>
    </INSDFeature_qual>
  </INSDFeature>
</INSDSeq_feature-table>

<INSDSeq_sequence>tagttcattgactaaggctccccattgactaaggcgactagcattgactaaggcaagc</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="37">
  <INSDSeq>
    <INSDSeq_length>16</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..16</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>modified_base</INSDFeature_key>
        <INSDFeature_location>6</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>N-(2-aminoethyl) glycine 5-nitroindole or N-(2-
aminoethyl) glycine 3-nitroindole</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>modified_base</INSDFeature_key>
        <INSDFeature_location>1..16</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>N-(2-aminoethyl) glycine nucleosides
(PNA)</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>cgccctnagtcaatggg</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="38">
  <INSDSeq>
    <INSDSeq_length>25</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..25</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>atcgggatcgcatattcgattggcc</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="39">
  <INSDSeq>
    <INSDSeq_length>25</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..25</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>ggccaatatggcttgcatccgat</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="40">
  <INSDSeq>
    <INSDSeq_length>20</INSDSeq_length>
    <INSDSeq_moltype>RNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..20</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>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</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>ccugucgt is attached at its 3' end to a linker which is attached to the 5' oxygen of the thymidine. The linker is (4-(3-hydroxybenzamido)butyl) phosphinic acid</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>modified_base</INSDFeature_key>
        <INSDFeature_location>1</INSDFeature_location>
        <INSDFeature_qual>

```



```
<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>19..114</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>translation</INSDQualifier_name>
      <INSDQualifier_value>MAPNTQTASPRALADSLMQLARQVSRLESGQ</INSDQualifier_value>
    </INSDQualifier>
    <INSDQualifier>
      <INSDQualifier_name>protein_id</INSDQualifier_name>
      <INSDQualifier_value>47</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>CDS</INSDFeature_key>
  <INSDFeature_location>115..222</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>translation</INSDQualifier_name>
      <INSDQualifier_value>MTMITDSLAVVLQRRDWENPGVTQLNRLAAHWCQK</INSDQualifier_value>
    </INSDQualifier>
    <INSDQualifier>
      <INSDQualifier_name>protein_id</INSDQualifier_name>
      <INSDQualifier_value>48</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>CDS</INSDFeature_key>
  <INSDFeature_location>251..283</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>translation</INSDQualifier_name>
      <INSDQualifier_value>MLRRQVNEVA</INSDQualifier_value>
    </INSDQualifier>
    <INSDQualifier>
      <INSDQualifier_name>protein_id</INSDQualifier_name>
      <INSDQualifier_value>49</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>caattcagggtggtgaatatggcgcccaatcgcgaaccgcctctcccgcgcttgccgattcattaatggaaagcgggcagt
gaatgaccatgattacggattcactggcgcgtcgttttacaacgtcgtgactgggaaaaccctggcgttacccaacttaatcgcttgcagcacattggtgtca
aaaataataataaccggatgtactatttaccctgatgctgcgtcgcagtggaatgaagtcgcttaagcaatcaatgctggatgcccgcgcagccttatccg
accaacatatcataa</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="47">
  <INSDSeq>
    <INSDSeq_length>31</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..31</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>
</SequenceData>
```



```
        </INSDQualifier>
      </INSDFeature_qual>
    </INSDFeature>
  </INSDSeq_feature-table>
  <INSDSeq_sequence>MAPNTQTASPRLADSLMQRLARQVSRLESGQ</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="48">
  <INSDSeq>
    <INSDSeq_length>35</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..35</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>MTMITDSLAVVLQRRDWENPGVTQLNRLAAHWCQK</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="49">
  <INSDSeq>
    <INSDSeq_length>10</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..10</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>MLRRQVNEVA</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="50">
  <INSDSeq>
    <INSDSeq_length>10</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..10</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>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
  </INSDSeq>
</SequenceData>
```

```

</INSDFeature>
<INSDFeature>
  <INSDFeature_key>SITE</INSDFeature_key>
  <INSDFeature_location>2</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>Homoserine</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>VARIANT</INSDFeature_key>
  <INSDFeature_location>3..6</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>A or G</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>AXXXXXYLGS</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="51">
  <INSDSeq>
    <INSDSeq_length>17</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..17</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>DISULFID</INSDFeature_key>
        <INSDFeature_location>join(4,15)</INSDFeature_location>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>CARBOHYD</INSDFeature_key>
        <INSDFeature_location>9</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>Asn side-chain N-linked to
asialyloligosaccharide</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>LEYCLKRWNETISHCAW</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="52">
  <INSDSeq>
    <INSDSeq_length>106</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..106</INSDFeature_location>
        <INSDFeature_qual>

```



```
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>VARIANT</INSDFeature_key>
  <INSDFeature_location>5..19</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>Up to 10 X residues may be
inserted</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>VARIANT</INSDFeature_key>
  <INSDFeature_location>21..39</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>Up to 10 X residues may be
inserted</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>KXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="57">
  <INSDSeq>
    <INSDSeq_length>6</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..6</INSDFeature_location>
        <INSDFeature_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>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>SITE</INSDFeature_key>
        <INSDFeature_location>1</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>This residue is linked N-terminally to a peptide
having an N-terminal Gly-Gly and a gap of undefined length</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>GGAGXX</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="58">
  <INSDSeq>
    <INSDSeq_length>45</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..45</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>1..45</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>translation</INSDQualifier_name>
      <INSDQualifier_value>MTGNKPETDVYEIL</INSDQualifier_value>
    </INSDQualifier>
    <INSDQualifier>
      <INSDQualifier_name>protein_id</INSDQualifier_name>
      <INSDQualifier_value>59</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>atgaccgaaataaacctgaaccgatgtttacgaaattttatga</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="59">
  <INSDSeq>
    <INSDSeq_length>14</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..14</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>MTGNKPETDVYEIL</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="60">
  <INSDSeq>
    <INSDSeq_length>22</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..22</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>organism</INSDQualifier_name>
            <INSDQualifier_value>Drosophila melanogaster</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>acattgaaatctcataccacttt</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
```

```
<SequenceData sequenceIDNumber="61">
  <INSDSeq>
    <INSDSeq_length>19</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..19</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>organism</INSDQualifier_name>
            <INSDQualifier_value>Drosophila virilis</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>mol_type</INSDQualifier_name>
            <INSDQualifier_value>genomic DNA</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>acatggatcccacgacttt</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="62">
  <INSDSeq>
    <INSDSeq_length>22</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..22</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>organism</INSDQualifier_name>
            <INSDQualifier_value>Drosophila simulans</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>mol_type</INSDQualifier_name>
            <INSDQualifier_value>genomic DNA</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>gtatggcgtcgtatsgtagttt</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="63">
  <INSDSeq>
    <INSDSeq_length>7</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..7</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>VARIANT</INSDFeature_key>
        <INSDFeature_location>7</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
    </INSDSeq_feature-table>
  </INSDSeq>
</SequenceData>
```

```
        <INSDQualifier_value>X is S or T</INSDQualifier_value>
      </INSDQualifier>
    </INSDFeature_qual>
  </INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>GGGJATX</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="64">
  <INSDSeq>
    <INSDSeq_length>7</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..7</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>VARIANT</INSDFeature_key>
        <INSDFeature_location>4</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>replace with I</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
      <INSDFeature>
        <INSDFeature_key>VARIANT</INSDFeature_key>
        <INSDFeature_location>7</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>replace with T</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>GGGLATS</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="65">
  <INSDSeq>
    <INSDSeq_length>7</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..7</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>GLPXRIC</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
```



```
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="66">
  <INSDSeq>
    <INSDSeq_length>7</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..7</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>GLPVVIC</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="67">
  <INSDSeq>
    <INSDSeq_length>7</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..7</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>GLPTRIC</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="68">
  <INSDSeq>
    <INSDSeq_length>7</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..7</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>GLPDRIC</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
```

```
<SequenceData sequenceIDNumber="69">
  <INSDSeq>
    <INSDSeq_length>36</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..36</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>organism</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>mol_type</INSDQualifier_name>
            <INSDQualifier_value>genomic DNA</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>cgaatgncctactacgaatgncacgaatgncccaca</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="70">
  <INSDSeq>
    <INSDSeq_length>36</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..36</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>organism</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>mol_type</INSDQualifier_name>
            <INSDQualifier_value>genomic DNA</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>cgaatgaccctactacgaatgbcacgaatgbccaca</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="71">
  <INSDSeq>
    <INSDSeq_length>36</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..36</INSDFeature_location>
        <INSDFeature_quals>
          <INSDQualifier>
            <INSDQualifier_name>organism</INSDQualifier_name>
            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>mol_type</INSDQualifier_name>
            <INSDQualifier_value>genomic DNA</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>cgaatgtcccactacgaatgvcacgaatgvccaca</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="72">
  <INSDSeq>
```

```
<INSDSeq_length>36</INSDSeq_length>
<INSDSeq_moltype>DNA</INSDSeq_moltype>
<INSDSeq_division>PAT</INSDSeq_division>
<INSDSeq_feature-table>
  <INSDFeature>
    <INSDFeature_key>source</INSDFeature_key>
    <INSDFeature_location>1..36</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>cgaatgcccactacgaatghcacgaatghcccaca</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="73">
  <INSDSeq>
    <INSDSeq_length>36</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..36</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>cgaatgcccactacgaatgdcacgaatgcccaca</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="74">
  <INSDSeq>
    <INSDSeq_length>400</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..400</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>join(1..79,142..212,272..400)</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>protein_id</INSDQualifier_name>
            <INSDQualifier_value>75</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
  </INSDSeq>
</SequenceData>
```

```
<INSDQualifier>
  <INSDQualifier_name>translation</INSDQualifier_name>
</INSDQualifier>
<INSDQualifier_value>MKTFAALLSAVTLALS SVRAQA AVWSQC GGTPGWTGETTCVAGSVCTSLSSSYSQCVPGSATSSAPAAPSATTSGPAPT DGTCSASGAWPPLT</INSDQualifier_value>
  </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>

<INSDSeq_sequence>atgaagacttttcgagccttgccttccgctgtcactctcgcgctctcggtgccgcccagggcggtgtctggagtcaatgtaagt
gccgctgcttttcattgatacgagactctacgccgagctgacgtgctaccgtataggtggcggtacaccgggttgacggggcagaccacttgctgtgctggt
tcggtttgtacctccttgagctcagtgagcagcttcaatccgtcgtcattgctcctcatgtattgacgattggccttcatagtcatactctcaatgcgttcc
gggctcgcgaactccagcgtcccgccggccccctcagcgacaacttcaggccccgcacctacggacggaacgtgctcggccagcggggcatggccgcattg
acctga</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="75">
  <INSDSeq>
    <INSDSeq_length>92</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..92</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>
</SequenceData>
<SequenceData sequenceIDNumber="76">
  <INSDSeq>
    <INSDSeq_length>9</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..9</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>AVLTYLRGE</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="77">
  <INSDSeq>
    <INSDSeq_length>9</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
```

```
<INSDFeature_key>SOURCE</INSDFeature_key>
<INSDFeature_location>1..9</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>AVLTYLRGA</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="78">
  <INSDSeq>
    <INSDSeq_length>9</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..9</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>AVPTYPRGE</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="79">
  <INSDSeq>
    <INSDSeq_length>9</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..9</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>AVAIGYRGE</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="80">
  <INSDSeq>
    <INSDSeq_length>8</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..8</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>AVLTYLGE</INSDSeq_sequence>  
</INSDSeq>  
</SequenceData>  
<SequenceData sequenceIDNumber="81">  
  <INSDSeq>  
    <INSDSeq_length>32</INSDSeq_length>  
    <INSDSeq_moltype>AA</INSDSeq_moltype>  
    <INSDSeq_division>PAT</INSDSeq_division>  
    <INSDSeq_feature-table>  
      <INSDFeature>  
        <INSDFeature_key>SOURCE</INSDFeature_key>  
        <INSDFeature_location>1..32</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>  
        </INSDFeature_qual>  
      </INSDFeature>  
      <INSDFeature>  
        <INSDFeature_key>VARIANT</INSDFeature_key>  
        <INSDFeature_location>4</INSDFeature_location>  
        <INSDFeature_qual>  
          <INSDQualifier>  
            <INSDQualifier_name>NOTE</INSDQualifier_name>  
            <INSDQualifier_value>N or T</INSDQualifier_value>  
          </INSDQualifier>  
        </INSDFeature_qual>  
      </INSDFeature>  
      <INSDFeature>  
        <INSDFeature_key>VARIANT</INSDFeature_key>  
        <INSDFeature_location>11</INSDFeature_location>  
        <INSDFeature_qual>  
          <INSDQualifier>  
            <INSDQualifier_name>NOTE</INSDQualifier_name>  
            <INSDQualifier_value>A or T</INSDQualifier_value>  
          </INSDQualifier>  
        </INSDFeature_qual>  
      </INSDFeature>  
      <INSDFeature>  
        <INSDFeature_key>VARIANT</INSDFeature_key>  
        <INSDFeature_location>12</INSDFeature_location>  
        <INSDFeature_qual>  
          <INSDQualifier>  
            <INSDQualifier_name>NOTE</INSDQualifier_name>  
            <INSDQualifier_value>K or Q</INSDQualifier_value>  
          </INSDQualifier>  
        </INSDFeature_qual>  
      </INSDFeature>  
      <INSDFeature>  
        <INSDFeature_key>VARIANT</INSDFeature_key>  
        <INSDFeature_location>18</INSDFeature_location>  
        <INSDFeature_qual>  
          <INSDQualifier>  
            <INSDQualifier_name>NOTE</INSDQualifier_name>  
            <INSDQualifier_value>K or Q or N or D</INSDQualifier_value>  
          </INSDQualifier>  
        </INSDFeature_qual>  
      </INSDFeature>  
    </INSDSeq_feature-table>  
  </INSDSeq>  
</SequenceData>
```

```
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>VARIANT</INSDFeature_key>
  <INSDFeature_location>21</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>R or G</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>VARIANT</INSDFeature_key>
  <INSDFeature_location>22</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>K or I or N</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>VARIANT</INSDFeature_key>
  <INSDFeature_location>24</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>L or V</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
<INSDFeature>
  <INSDFeature_key>VARIANT</INSDFeature_key>
  <INSDFeature_location>29</INSDFeature_location>
  <INSDFeature_qual>
    <INSDQualifier>
      <INSDQualifier_name>NOTE</INSDQualifier_name>
      <INSDQualifier_value>M or L</INSDQualifier_value>
    </INSDQualifier>
  </INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>LEGXEQFINAXXIIRHPXYBXXTXBNDIXLIK</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="82">
  <INSDSeq>
    <INSDSeq_length>32</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..32</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>LEGNEQFINAAKIIIRHPQYDRKTLNNDIMLIK</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="83">
  <INSDSeq>
    <INSDSeq_length>32</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
```

```

<INSDSeq_feature-table>
  <INSDFeature>
    <INSDFeature_key>SOURCE</INSDFeature_key>
    <INSDFeature_location>1..32</INSDFeature_location>
    <INSDFeature_qual>
      <INSDQualifier>
        <INSDQualifier_name>ORGANISM</INSDQualifier_name>
        <INSDQualifier_value>Pongo abelii</INSDQualifier_value>
      </INSDQualifier>
      <INSDQualifier>
        <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
        <INSDQualifier_value>protein</INSDQualifier_value>
      </INSDQualifier>
    </INSDFeature_qual>
  </INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>LEGNEQFINAAKIRHPQYDRKTVNNDIMLIK</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="84">
  <INSDSeq>
    <INSDSeq_length>32</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..32</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Papio anubis</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>LEGTEQFINAAKIRHPDYDRKTLNNDILLIK</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="85">
  <INSDSeq>
    <INSDSeq_length>32</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..32</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Rhinopithecus roxellana</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>LEGTEQFINAAKIRHPNYRITLNDNDILLIK</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="86">
  <INSDSeq>
    <INSDSeq_length>32</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>

```



```
<INSDFeature_key>SOURCE</INSDFeature_key>
<INSDFeature_location>1..32</INSDFeature_location>
<INSDFeature_qual>
  <INSDQualifier>
    <INSDQualifier_name>ORGANISM</INSDQualifier_name>
    <INSDQualifier_value>Pan paniscus</INSDQualifier_value>
  </INSDQualifier>
  <INSDQualifier>
    <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
    <INSDQualifier_value>protein</INSDQualifier_value>
  </INSDQualifier>
</INSDFeature_qual>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>LEGNEQFINAAKIIIRHPKYNRITLNDIMLIK</INSDSeq_sequence>
</INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="87">
  <INSDSeq>
    <INSDSeq_length>32</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..32</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Rhinopithecus bieti</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>LEGNEQFINATKIIIRHPKYNGNTLNDIMLIK</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="88">
  <INSDSeq>
    <INSDSeq_length>32</INSDSeq_length>
    <INSDSeq_moltype>AA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..32</INSDFeature_location>
        <INSDFeature_qual>
          <INSDQualifier>
            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
            <INSDQualifier_value>Rhinopithecus roxellana</INSDQualifier_value>
          </INSDQualifier>
          <INSDQualifier>
            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
            <INSDQualifier_value>protein</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_qual>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>LEGNEQFINATQIIIRHPKYNGNTLNDIMLIK</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
</ST26SequenceListing>
```

[End of Annex VI to ST.26 and of Standard]

[Fin del Anexo II y del documento]