**NORMA ST.26**

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

Proyecto final

*Propuesta presentada por el Equipo Técnico SEQL para examen y aprobación por el CWS/4*

ÍNDICE

INTRODUCCIÓN 2

DEFINICIONES 2

ÁMBITO 3

REFERENCIAS 3

PRESENTACIÓN DE SECUENCIAS 3

*Secuencias de nucleótidos* 3

*Secuencias de aminoácidos* 5

*Presentación de casos especiales* 7

ESTRUCTURA DE LA LISTA DE SECUENCIAS EN FORMATO XML 7

*Elemento raíz* 8

*Parte de información general* 9

*Parte de datos de secuencia* 12

*Cuadro de características* 14

*Claves de caracterización* 14

*Claves de caracterización obligatorias* 14

*Localización de característica* 14

*Calificadores de caracterización* 16

*Calificadores de caracterización obligatorios* 17

*Elementos de los calificadores* 17

*Texto libre* 19

*Secuencias codificadoras* 19

*Variantes* 19

**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 de la tabla 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)

**NORMA ST.26**

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

Proyecto final

*Propuesta presentada por el Equipo Técnico SEQL para examen y aprobación por el CWS/4*

## 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 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:
3. 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;
4. 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;
5. facilitar la búsqueda de datos en las secuencias; y
6. permitir el intercambio electrónico de datos sobre las secuencias y la introducción de esos datos en bases de datos informatizadas.

## DEFINICIONES

1. A los efectos de la presente Norma:
2. 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 que se trata de aminoácidos modificados tal como se prevé en el párrafo 29.
3. 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.
4. 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.
5. por “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, una base modificada o sintética de purina o pirimidina, o una ribosa o deoxirribosa 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.
6. por “residuo” se entenderá todo nucleótido o aminoácido individual de una secuencia.
7. por “identificador de secuencia” se entenderá un número entero único que se asigna a cada secuencia de la lista de secuencias.
8. 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 la(s) secuencia(s) divulgada(s) de nucleótidos y/o aminoácidos junto con toda otra descripción detallada.
9. 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.
10. por nucleótido o aminoácido “desconocido” se entenderá la presencia de un único nucleótido o aminoácido cuya identidad es desconocida o no se divulga.

## ÁMBITO

1. 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.
2. 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.
3. 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:

a) una secuencia no ramificada o una porción 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

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

1. Una lista de secuencias no deberá incluir ninguna secuencia que tenga menos de diez nucleótidos específicamente definidos, o menos de cuatro aminoácidos específicamente definidos.

## REFERENCIAS

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

*International Nucleotide Sequence   
Database Collaboration* (INSDC) <http://www.insdc.org/>;

Norma ISO 639-1 - *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](http://www.wipo.int/standards/es/pdf/03-02-01.pdf) Forma normalizada de designar las fechas según el calendario gregoriano;

Norma técnica de la OMPI [ST.3](http://www.wipo.int/standards/es/pdf/03-03-01.pdf) Códigos normalizados de dos letras, recomendados para la representación de Estados, otras entidades y organizaciones intergubernamentales;

Norma técnica de la OMPI [ST.16](http://www.wipo.int/standards/es/pdf/03-16-01.pdf) Código normalizado para la identificación de los diferentes tipos de documentos de patente;

Norma técnica de la OMPI [ST.25](http://www.wipo.int/standards/es/pdf/03-25-01.pdf) Norma para la presentación de listas de secuencias de nucleótidos y aminoácidos en solicitudes de patente.

## PRESENTACIÓN DE SECUENCIAS

1. Se asignará a cada secuencia un identificador de secuencia distinto. Los identificadores de secuencias comenzarán 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 utilizará 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*

1. Toda secuencia de nucleótidos sólo deberá representarse mediante una cadena única, en el sentido 5’ a 3’ y de izquierda a derecha. Los valores 5’ y 3’ no deberán aparecer en la secuencia. Toda secuencia de nucleótidos de doble cadena divulgada mediante la enumeración de los residuos de ambas cadenas deberá presentarse de la siguiente manera:

a) una única secuencia 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.

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’.
2. 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.
3. 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 utilizarán únicamente letras minúsculas. Todo símbolo utilizado para representar un nucleótido equivale a un único residuo.
4. 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 tal como se prevé en el párrafo 18.
5. Si fuera necesario utilizar un símbolo de ambigüedad (que represente dos o más bases alternativas), deberá utilizarse el símbolo más restrictivo. Por ejemplo, si una base en una determinada posición pudiera ser “a” o “g”, se utilizará “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 20. El símbolo “n” no podrá 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 tal como se prevé en los párrafos 16 y 17 o 20.
6. Los nucleótidos modificados deberán representarse en la secuencia como las correspondientes bases no modificadas, 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), tales como los nucleótidos que no existen 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.
7. Los nucleótidos modificados deberán describirse con detalle en el cuadro de características (véanse los párrafos 59 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 de la base modificada deberá indicarse como valor en un calificador “note”. 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.
8. 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.
9. Los siguientes ejemplos ilustran la presentación de los nucleótidos modificados en la forma prevista en los párrafos 16 y 17 *supra*:

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\_quals>

<INSDQualifier>

<INSDQualifier\_name>mod\_base</INSDQualifier\_name>

<INSDQualifier\_value>i</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

Ejemplo 2: Nucleótido modificado 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\_quals>

<INSDQualifier>

<INSDQualifier\_name>mod\_base</INSDQualifier\_name>

<INSDQualifier\_value>OTHER</INSDQualifier\_value>

</INSDQualifier>

<INSDQualifier>

<INSDQualifier\_name>note</INSDQualifier\_name>

<INSDQualifier\_value>xanthine</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

1. Todo nucleótido “desconocido” deberá representarse mediante el símbolo “n” en la secuencia. Los nucleótidos “desconocidos” deberán 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.
2. 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 la sintaxis “x..y” como descriptor de localización en el elemento INSDFeature\_location (véanse los párrafos 65 a 72). Para la presentación de las variantes de secuencia, es decir, supresiones, inserciones o sustituciones, véanse los párrafos 92 a 97.
3. El siguiente ejemplo ilustra la presentación de una región de nucleótidos modificados para los cuales se aplica la misma descripción en la forma prevista en el párrafo 21 *supra*:

<INSDFeature>

<INSDFeature\_key>modified\_base</INSDFeature\_key>

<INSDFeature\_location>358..485</INSDFeature\_location>

<INSDFeature\_quals>

<INSDQualifier>

<INSDQualifier\_name>mod\_base</INSDQualifier\_name>

<INSDQualifier\_value>OTHER</INSDQualifier\_value>

</INSDQualifier>

<INSDQualifier>

<INSDQualifier\_name>note</INSDQualifier\_name>

<INSDQualifier\_value>isoguanine</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

### *Secuencias de aminoácidos*

1. Los aminoácidos de una secuencia proteínica o peptídica deberán relacionarse 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.
2. La numeración de las posiciones de aminoácidos deberá comenzar en el primer aminoácido de la secuencia, con el 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 a lo largo de toda la secuencia en el sentido de amino a carboxilo.
3. 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 utilizarán únicamente letras mayúsculas. Todo símbolo utilizado para representar un aminoácido equivale a un único residuo.
4. Si fuese necesario utilizar un símbolo de ambigüedad (que represente dos o más aminoácidos posibles), deberá utilizarse el símbolo más restrictivo. Por ejemplo, si un aminoácido en una posición dada podría ser un ácido aspártico o asparagina, deberá 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 tal como se prevé en los párrafos 28 a 30 o 31 a 33. El símbolo “X” no podrá utilizarse para representar un elemento distinto a un aminoácido. Un único aminoácido podrá representarse mediante el símbolo “X”, junto con una descripción detallada tal como se prevé en los párrafos  28 a 30 o 31 a 33. Para la presentación de las variantes de secuencia, es decir, supresiones, inserciones, o sustituciones, véanse los párrafos 92 a 97.
5. Las secuencias de aminoácidos separadas por uno o más espacios en blanco o símbolos internos de terminación, por ejemplo, “Ter” o asterisco “\*” o punto “.”, en una divulgación, deberán presentarse 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 6. Cada secuencia distinta de este tipo deberá presentarse en la lista de secuencias con 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 símbolos de terminación ni espacios en las secuencias presentadas en una lista de secuencias.
6. Los aminoácidos modificados, incluidos los D-aminoácidos, deberán 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), deberá representarse por el símbolo “X”. El símbolo “X” equivale a un único residuo.
7. Los aminoácidos modificados deberán describirse con detalle en el cuadro de características (véanse los párrafos 60 y siguientes). La clave de caracterización “MOD\_RES” deberá utilizarse para las modificaciones postraduccionales de los aminoácidos junto con el calificador “NOTE” y la clave de caracterización “SITE” para otros aminoácidos modificados 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.
8. Los siguientes ejemplos ilustran la presentación de aminoácidos modificados en la forma prevista en el párrafo 29 *supra*:

Ejemplo 1: Aminoácido modificado (modificación postraduccional)

<INSDFeature>

<INSDFeature\_key>MOD\_RES</INSDFeature\_key>

<INSDFeature\_location>3</INSDFeature\_location>

<INSDFeature\_quals>

<INSDQualifier>

<INSDQualifier\_name>NOTE</INSDQualifier\_name>

<INSDQualifier\_value>3-Hyp</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

Ejemplo 2: Aminoácido modificado (modificación no postraduccional)

<INSDFeature>

<INSDFeature\_key>SITE</INSDFeature\_key>

<INSDFeature\_location>3</INSDFeature\_location>

<INSDFeature\_quals>

<INSDQualifier>

<INSDQualifier\_name>NOTE</INSDQualifier\_name>

<INSDQualifier\_value>Orn</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

Ejemplo 3: D-aminoácido

<INSDFeature>

<INSDFeature\_key>SITE</INSDFeature\_key>

<INSDFeature\_location>9</INSDFeature\_location>

<INSDFeature\_quals>

<INSDQualifier>

<INSDQualifier\_name>NOTE</INSDQualifier\_name>

<INSDQualifier\_value>D-Arginine</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

1. 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.
2. 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”.
3. Los siguientes ejemplos ilustran la presentación de aminoácidos “desconocidos” u “otros” en la forma prevista en los párrafos 31 y 32 *supra*:

Ejemplo 1: Aminoácido “desconocido”

<INSDFeature>

<INSDFeature\_key>UNSURE</INSDFeature\_key>

<INSDFeature\_location>3</INSDFeature\_location>

<INSDFeature\_quals>

<INSDQualifier>

<INSDQualifier\_name>NOTE</INSDQualifier\_name>

<INSDQualifier\_value>A or V</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

Ejemplo 2: Aminoácido “otro”

<INSDFeature>

<INSDFeature\_key>SITE</INSDFeature\_key>

<INSDFeature\_location>3</INSDFeature\_location>

    <INSDFeature\_quals>

     <INSDQualifier>

         <INSDQualifier\_name>NOTE</INSDQualifier\_name>

            <INSDQualifier\_value>Homoserine</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

1. 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..y” como descriptor de localización en el elemento INSDFeature\_location (véanse los párrafos 65 a 71). Para la presentación de variantes de secuencia, a saber, supresiones, inserciones, o sustituciones, véanse los párrafos 92 a 97.

### *Presentación de casos especiales*

1. 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 identificador de secuencia.
2. Toda secuencia divulgada mediante la enumeración de sus residuos que contiene regiones de residuos específicamente enumerados separados por una o más regiones de residuos “n” o “X” contiguos (véanse los párrafos 15 y 26, respectivamente), en la que el número exacto de residuos de cada región se divulga, deberá figurar en la lista de secuencias de una única secuencia con un único identificador de secuencia.
3. Toda secuencia divulgada mediante la enumeración de sus residuos que contiene regiones de residuos específicamente enumerados y separados por uno o más huecos compuestos por un número desconocido o no divulgado de residuos 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 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

1. Según lo previsto en el párrafo 5 *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.

1. 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”>.

1. 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;

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

* + deberá estar compuesta por caracteres imprimibles (incluido el carácter de espacio) de la tabla 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
  + las únicas entidades de caracteres permitidas son las entidades predefinidas descritas en el párrafo 41.

1. 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** |
| < | &lt; |
| > | &gt; |
| & | &amp; |
| “ | &quot; |
| ' | &apos; |

Véase el párrafo 72 para consultar un ejemplo.

1. 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 deberán aparecer en la instancia XML.

### *Elemento raíz*

1. 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 |

1. 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\_SEQL.xml” softwareName=“SEQL-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*

1. 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** |
| --- | --- | --- |
| ApplicationIdentification  La ApplicationIdentification se compone de: | La identificación de la solicitud para la cual se presenta la lista de secuencias | Obligatorio cuando una lista de secuencias se suministra en cualquier momento posterior a la asignación del número de solicitud |
| IPOfficeCode | El código ST.3 de la oficina de presentación | Obligatorio |
| ApplicationNumberText | La identificación de la solicitud suministrada por la oficina de presentación (por ejemplo, PCT/IB2013/099999) | Obligatorio |
| FilingDate | 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) | Obligatorio cuando una lista de secuencias se suministra en cualquier momento posterior a la asignación de una fecha de presentación |
| ApplicantFileReference | Un identificador único asignado por el solicitante para identificar una solicitud específica, escrito en los caracteres descritos en el párrafo 40.b) | Obligatorio cuando una lista de secuencias se suministra en cualquier momento anterior a la asignación del número de solicitud; de los contrario, Facultativo |
| EarliestPriorityApplicationIdentification | La identificación de la reivindicación de prioridad más antigua (también contiene IPOfficeCode, ApplicationNumberText y FilingDate, véase ApplicationIdentification *supra*) | Obligatorio cuando se reivindica la prioridad |
| ApplicantName | 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. | Obligatorio |
| ApplicantNameLatin | 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) | Obligatorio cuando ApplicantName contiene caracteres no latinos |
| InventorName | 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. | Facultativo |
| InventorNameLatin | Si InventorName se escribe en caracteres distintos a los descritos en el párrafo 40.b), deberá 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 |

1. 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\_SEQL.xml” softwareName=“SEQL-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\_SEQL.xml” softwareName=“SEQL-software-name” softwareVersion=“1.0” productionDate=“2015-05-10”>

<ApplicationIdentification>

<IPOfficeCode>US</IPOfficeCode>

<ApplicationNumberText>14/999,999</ApplicationNumberText>

<FilingDate>2015-01-05</FilingDate>

</ApplicationIdentification>

<ApplicantFileReference>AB123</ApplicantFileReference>

<EarliestPriorityApplicationIdentification>

<IPOfficeCode>IB</IPOfficeCode>

<ApplicationNumberText>PCT/IB2014/099999</ApplicationNumberText>

<FilingDate>2014-07-10</FilingDate>

</EarliestPriorityApplicationIdentification>

<ApplicantName languageCode=“EN”>GENOS Co., Inc.</ApplicantName>

<InventorName languageCode=“EN”>Keiko Nakamura</InventorName>

<InventionTitle languageCode=“EN”>SIGNAL RECOGNITION PARTICLE RNA AND PROTEINS</InventionTitle>

<SequenceTotalQuantity>9</SequenceTotalQuantity>

<SequenceData sequenceIDNumber=“1”> {...}\* </SequenceData>

<SequenceData sequenceIDNumber=“2”> {...} </SequenceData>

<SequenceData sequenceIDNumber=“3”> {...} </SequenceData>

<SequenceData sequenceIDNumber=“4”> {...} </SequenceData>

<SequenceData sequenceIDNumber=“5”> {...} </SequenceData>

<SequenceData sequenceIDNumber=“6”> {...} </SequenceData>

<SequenceData sequenceIDNumber=“7”> {...} </SequenceData>

<SequenceData sequenceIDNumber=“8”> {...} </SequenceData>

<SequenceData sequenceIDNumber=“9”> {...} </SequenceData>

</ST26SequenceListing>

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

1. 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 el párrafo 8.b) 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 deberá indicarse una transliteración o traducción del nombre del inventor en caracteres del alfabeto latino en el elemento InventorNameLatin.
2. 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 el párrafo 8.b) deberá indicarse en el atributo languageCode del elemento.
3. 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="JA">出願製薬株式会社</ApplicantName>

<ApplicantNameLatin>Shutsugan Pharmaceuticals Kabushiki Kaisha</ApplicantNameLatin>

<InventorName languageCode ="JA">特許　太郎</InventorName>

<InventorNameLatin>Taro Tokkyo</InventorNameLatin>

<InventionTitle languageCode="JA"> efgタンパク質のためのマウスabcd-1遺伝子</InventionTitle>

<InventionTitle languageCode="EN"> Mus musculus abcd-1 gene for efg protein </InventionTitle>

<InventionTitle languageCode="FR"> Gène abcd-1 de Mus musculus pour protéine efg </InventionTitle>

### *Parte de datos de secuencia*

1. La parte de datos de secuencia deberá constar de uno o más elementos SequenceData, y cada elemento contendrá información sobre una sola secuencia.
2. Cada elemento SequenceData deberá tener un atributo obligatorio sequenceIDNumber, el cual contiene el identificador de secuencia (véase el párrafo 9) de cada secuencia. Por ejemplo:

<SequenceData sequenceIDNumber=“1”>

1. 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** |
| INSDSeq\_length | Longitud de la secuencia | Obligatorio | Obligatorio, sin ningún valor |
| INSDSeq\_moltype | Tipo de molécula | Obligatorio | Obligatorio, sin ningún valor |
| INSDSeq\_division | Indicación de que una secuencia está asociadas a una solicitud de patente | Obligatorio con el valor “PAT” | Obligatorio, sin ningún valor |
| INSDSeq\_feature-table | Lista de anotaciones de la secuencia | Obligatorio | NO debe incluirse |
| INSDSeq\_sequence | Secuencia | Obligatorio | Obligatorio con el valor “000” |

1. 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>

1. El elemento INSDSeq\_moltype deberá divulgar el tipo de molécula que se está presentando. En el caso de las secuencias de nucleótidos, el tipo de molécula deberá indicarse como ADN o ARN. Para las secuencias proteínicas o polipéptidicas, 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 85). Por ejemplo:

<INSDSeq\_moltype>AA</INSDSeq\_moltype>

1. 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 deberá 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.
2. 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\_quals>

<INSDQualifier>

<INSDQualifier\_name>organism</INSDQualifier\_name>

<INSDQualifier\_value>synthetic construct</INSDQualifier\_value>

</INSDQualifier>

<INSDQualifier>

<INSDQualifier\_name>mol\_type</INSDQualifier\_name>

<INSDQualifier\_value>other DNA</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

<INSDFeature>

<INSDFeature\_key>misc\_feature</INSDFeature\_key>

<INSDFeature\_location>1..60</INSDFeature\_location>

<INSDFeature\_quals>

<INSDQualifier>

<INSDQualifier\_name>note</INSDQualifier\_name>

<INSDQualifier\_value>DNA fragment</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

<INSDFeature>

<INSDFeature\_key>misc\_feature</INSDFeature\_key>

<INSDFeature\_location>61..120</INSDFeature\_location>

<INSDFeature\_quals>

<INSDQualifier>

<INSDQualifier\_name>note</INSDQualifier\_name>

<INSDQualifier\_value>RNA fragment</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

</INSDSeq\_feature-table>

<INSDSeq\_sequence> cgacccacgcgtccgaggaaccaaccatcacgtttgaggacttcgtgaaggaattggataatacccgtccctaccaaaatggcgagcgccgactcattgctcctcgtaccgtcgagcggc

</INSDSeq\_sequence>

</INSDSeq>

1. El elemento INSDSeq\_sequence deberá divulgar la secuencia. Los residuos de la secuencia deberán presentarse de forma contigua utilizando únicamente 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 números, signos de puntuación o caracteres en blanco.
2. Toda secuencia ignorada deliberadamente deberá presentarse 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”.

1. El siguiente ejemplo ilustra la presentación 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*

1. 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.
2. Cada elemento INSDFeature describe una característica, y consiste en elementos dependientes de la siguiente manera:

|  |  |  |
| --- | --- | --- |
| **Elemento** | **Descripción** | **Obligatorio/Facultativo** |
| INSDFeature\_key | Palabra o abreviatura que indica una característica | Obligatorio |
| INSDFeature\_location | Región de la secuencia presentada que corresponde a la característica | Obligatorio |
| INSDFeature\_quals | Calificador que contiene información auxiliar acerca de la característica | Obligatorio si la clave de caracterización requiere uno o más calificadores, por ejemplo, source; de lo contrario, Facultativo |

### *Claves de caracterización*

1. 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*

1. 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 deberán 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.
2. Algunas claves de caracterización requieren otra clave 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*

1. 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 68 a 71).
2. 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 68 a 71). El descriptor de localización no deberá incluir números de residuos fuera de la serie de la secuencia indicada en el elemento INSDSeq\_sequence.
3. 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 | X | Designa un único residuo en la secuencia presentada. |
| Números de residuos que limitan un tramo de secuencia | x..y | 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 | <x  >x  <x..y  x..>y | Designa una región que incluye un residuo o una serie de residuos especificada y que se extiende más allá del residuo especificado. Los símbolos '<' y '>' 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 las características se extienden más allá del número de residuo especificado. |
| 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. |

1. 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** |
| join(location,location, ... location) | Las localizaciones indicadas están unidas (colocadas extremo con extremo) para formar una secuencia contigua. |
| order(location,location, ... location) | 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** |
| complement(location) | 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’. |

1. 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 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).
2. 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.
3. Los siguientes ejemplos ilustran localizaciones de características en la forma prevista en los párrafos 65 a 70 *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. |
| join(12..78,134..202) | Indica que las regiones 12 a 78 y 134 a 202 deben unirse para formar una secuencia contigua. |

b) localizaciones únicamente para nucleótidos:

|  |  |
| --- | --- |
| **Ejemplo de localización** | **Descripción** |
| complement(34..126) | Comienza en la base complementaria a 126 y termina en la base complementaria a la base 34 (la característica está en la cadena complementaria a la cadena presentada). |
| complement(join(2691..4571, 4918..5163)) | Une las bases de 2691 a 4571 y 4918 a 5163, y luego complementa los segmentos unidos (la característica está en la cadena complementaria a la cadena presentada). |
| join(complement(4918..5163), complement(2691..4571)) | 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). |

1. En una instancia XML de una lista de secuencias, los caracteres “<” y “>” de un descriptor de localización deberán sustituirse por las entidades predefinas 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*

1. 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 86 y 87);

b) vocabulario controlado o valores enumerados (por ejemplo, un número o fecha); y

c) secuencias.

1. 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.
2. Toda secuencia contemplada en el párrafo 6 que se indique como valor calificador deberá presentarse de manera clara en la lista de secuencias con su propio identificador de secuencia.

### *Calificadores de caracterización obligatorios*

1. 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*

1. El elemento INSDFeature\_quals 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** |
| INSDQualifier\_name | Nombre del calificador (véanse las Secciones 6 y 8 del Anexo I) | Obligatorio |
| INSDQualifier\_value | 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) |

1. 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 deberán seleccionarse a partir de una base de datos taxonómica.
2. Si la secuencia existe en estado natural y el organismo fuente tiene una designación de género y especie en latín, esa designación deberá 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.
3. Los siguientes ejemplos ilustran la fuente de las secuencias presentadas en la forma prevista en los párrafos 78 y 79 *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\_quals>

<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\_quals>

</INSDFeature>

</INSDSeq\_feature-table>

Ejemplo 2: Fuente de una secuencia proteínica

<INSDSeq\_feature-table>

<INSDFeature>

<INSDFeature\_key>SOURCE</INSDFeature\_key>

<INSDFeature\_location>1..174</INSDFeature\_location>

<INSDFeature\_quals>

<INSDQualifier>

<INSDQualifier\_name>ORGANISM</INSDQualifier\_name>

<INSDQualifier\_value>Homo sapiens</INSDQualifier\_value>

</INSDQualifier>

<INSDQualifier>

<INSDQualifier\_name>MOL\_TYPE</INSDQualifier\_name>

<INSDQualifier\_value>protein</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

</INSDSeq\_feature-table>

1. 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>

1. Si la fuente de la secuencia es 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>

1. 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>

1. 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\_quals>

<INSDQualifier>

<INSDQualifier\_name>ORGANISM</INSDQualifier\_name>

<INSDQualifier\_value>synthetic construct</INSDQualifier\_value>

</INSDQualifier>

<INSDQualifier>

<INSDQualifier\_name>MOL\_TYPE</INSDQualifier\_name>

<INSDQualifier\_value>protein</INSDQualifier\_value>

</INSDQualifier>

<INSDQualifier>

<INSDQualifier\_name>NOTE</INSDQualifier\_name>

<INSDQualifier\_value>synthetic peptide used as assay for antibodies</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

</INSDSeq\_feature-table>

1. 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*

1. 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 preferentemente en inglés.
2. 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*

1. 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 deberá identificar la localización de la característica “CDS” y debe incluir el codón de terminación.
2. 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 de la tabla de códigos normalizados (véase el Cuadro 5 de la Sección 9 del Anexo I).
3. Toda secuencia proteínica codificada por la secuencia codificadora y divulgada en un calificador “translation” contemplado en el párrafo 6 deberá tener su propio identificador de secuencia, y presentarse en la lista de secuencias. El identificador de secuencia asignado a la secuencia proteínica deberá figurar como valor del calificador “protein\_id” con la clave caracterización “CDS”. El calificador “ORGANISM” de la clave de caracterización “SOURCE” para la secuencia proteínica 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\_quals>

<INSDQualifier>

<INSDQualifier\_name>transl\_table</INSDQualifier\_name>

<INSDQualifier\_value>11</INSDQualifier\_value>

</INSDQualifier>

<INSDQualifier>

<INSDQualifier\_name>translation</INSDQualifier\_name>

<INSDQualifier\_value>

MLVHLERTTIMFDFSSLINLPLIWGLLIAIAVLLYILMDGFDLGIGILLPFAPSDKCRDHMISSIAPFWDGNETWLVLGGGGLFAAFPLAYSILMPAFYIPIIIMLLGLIVRGVSFEFRFKAEGKYRRLWDYAFHFGSLGAAFCQGMILGAFIHGVEVNGRNFSGGQLM

</INSDQualifier\_value>

</INSDQualifier>

<INSDQualifier>

<INSDQualifier\_name>protein\_id</INSDQualifier\_name>

<INSDQualifier\_value>89</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

</INSDSeq\_feature-table>

### *Variantes*

1. 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 6, deberá presentarse en la lista de secuencias con su propio identificador de secuencia.
2. 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 presentarse en la lista de secuencias. Cuando se indica en la lista de secuencias, tal variante de secuencia:

a) podrá presentarse 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) deberá presentarse como una secuencia distinta con su propio identificador de secuencia, si contiene variaciones en varias localizaciones distintas y esas variaciones aparecen de forma interdependiente; y

c) debe presentarse como una secuencia distinta con 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).

1. 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 | 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 | 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. |

1. 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. Toda supresión deberá representarse mediante un valor calificador vacío. Todo residuo insertado o sustituido deberá indicarse en el calificador “replace” o “NOTE”. El valor de los calificadores “replace” y “NOTE” es un texto libre que no deberá ser superior a 1000 caracteres, tal como se prevé en el párrafo 87. Para las secuencias contempladas en el párrafo 6 que se presentan como inserción o sustitución en un valor calificador, véase el párrafo 97. El valor calificador puede contener una lista de los residuos alternativos que pueden insertarse o sustituirse.
2. Los símbolos descritos en el Anexo I (véanse los Cuadros 1 a 4 de las Secciones 1 a 4, respectivamente) deberán utilizarse para representar las variantes de residuos, dado el caso. 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.
3. Los siguientes ejemplos ilustran la presentación de variantes en la forma prevista en los párrafos 92 a 95 *supra*:

Ejemplo 1: Clave de caracterización “variation” para una sustitución en una secuencia de nucleótidos.

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

<INSDFeature>

<INSDFeature\_key>variation</INSDFeature\_key>

<INSDFeature\_location>413</INSDFeature\_location>

<INSDFeature\_quals>

<INSDQualifier>

<INSDQualifier\_name>replace</INSDQualifier\_name>

<INSDQualifier\_value>c</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</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\_quals>

<INSDQualifier>

<INSDQualifier\_name>replace</INSDQualifier\_name>

<INSDQualifier\_value></INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</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\_quals>

<INSDQualifier>

<INSDQualifier\_name>replace</INSDQualifier\_name>

<INSDQualifier\_value>atgccaaatat</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

Ejemplo 4: 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 I, A, F, Y, aIle, MeIle, o Nle.

<INSDFeature>

<INSDFeature\_key>VARIANT</INSDFeature\_key>

<INSDFeature\_location>100</INSDFeature\_location>

<INSDFeature\_quals>

<INSDQualifier>

<INSDQualifier\_name>NOTE</INSDQualifier\_name>

<INSDQualifier\_value>I, A, F, Y, aIle, MeIle, or Nle </INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</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 or His.

<INSDFeature>

<INSDFeature\_key>VARIANT</INSDFeature\_key>

<INSDFeature\_location>100</INSDFeature\_location>

<INSDFeature\_quals>

<INSDQualifier>

<INSDQualifier\_name>NOTE</INSDQualifier\_name>

<INSDQualifier\_value>not K, R, or H</INSDQualifier\_value>

</INSDQualifier>

</INSDFeature\_quals>

</INSDFeature>

1. Toda secuencia contemplada en el párrafo 6 que se indique como una inserción o sustitución en un valor calificador para una anotación de secuencia principal también deberá presentarse en la lista de secuencias con su propio identificador de secuencia.

[Sigue el Anexo I a la ST.26]