**NORME ST.26**

recommandation de NORME RELATIVE À LA PRÉSENTATION DES LISTAGES DES SÉQUENCES DE NUCLÉOTIDES ET D’ACIDES AMINÉS EN LANGAGE XML (*EXTENSIBLE MARKUP LANGUAGE*)

Version finale

*Proposition présentée par l’équipe d’experts SEQL pour examen et adoption par le CWS à sa quatrième session*

TABLE DES MATIÈRES

INTRODUCTION 2

DÉFINITIONS 2

PORTÉE 3

RÉFÉRENCES 3

PRÉSENTATION DES SÉQUENCES 3

*Séquences de nucléotides* 3

*Séquences d’acides aminés* 5

*Présentation de cas particuliers* 7

STRUCTURE DU LISTAGE DE SÉQUENCES EN XML 7

*Élément racine* 8

*Partie consacrée aux informations générales* 8

*Partie consacrée aux données sur les séquences* 11

*Tableau de caractéristiques* 13

*Clés de caractérisation* 14

*Clés de caractérisation obligatoires* 14

*Emplacement de la caractéristique* 14

*Qualificateurs de caractéristiques* 16

*Qualificateurs de caractéristiques obligatoires* 16

*Éléments des qualificateurs* 16

*Texte libre* 18

*Séquences de codage* 18

*Variantes* 19

**ANNEXES**

Annexe I – Vocabulaire contrôlé

Annexe II – Définition de type de document (DTD) pour le listage des séquences

Annexe III – Exemple de listage des séquences (fichier XML)

Annexe IV – Sous‑ensemble de caractères provenant du tableau de codes des caractères latins de base de la norme Unicode

Annexe V – Prescriptions supplémentaires en matière d’échange de données (uniquement pour les offices de brevets)

**NORME ST.26**

recommandation de NORME RELATIVE À LA PRÉSENTATION DES LISTAGES DES SÉQUENCES DE NUCLÉOTIDES ET D’ACIDES AMINÉS EN LANGAGE XML (*EXTENSIBLE MARKUP LANGUAGE*)

Version finale

*Proposition présentée par l’équipe d’experts SEQL pour examen et adoption par le CWS à sa quatrième session*

## INTRODUCTION

1. La présente norme définit la manière dont des séquences de nucléotides et d’acides aminés doivent être divulguées dans une demande de brevet pour pouvoir être jointes à un listage des séquences. Elle précise les caractéristiques de ces divulgations et la définition de type de document (DTD) à employer lorsque le listage des séquences est effectué au format XML (*eXtensible Markup Language*). Il est recommandé que les offices de propriété industrielle acceptent tous les listages de séquences conformes à cette norme qui sont déposés en tant que partie intégrante d’une demande de brevet ou en relation avec une demande de brevet.
2. Cette norme a pour but :
3. de permettre aux déposants d’établir, dans le cadre d’une demande de brevet, un listage des séquences unique qui soit acceptable pour les procédures internationales et nationales ou régionales;
4. d’accroître la précision et la qualité de la présentation des séquences pour faciliter leur diffusion dans l’intérêt des déposants, du public et des examinateurs;
5. de faciliter la recherche de données sur ces séquences; et
6. de permettre l’échange de données sur les séquences sous forme électronique et l’incorporation de ces données dans des bases de données informatisées.

## DÉFINITIONS

1. Aux fins de la présente norme, l’expression :
2. “acide aminé” désigne tout acide aminé pouvant être représenté à l’aide des symboles indiqués dans le tableau I (voir section 3, tableau 3). Ces acides aminés comprennent notamment les D‑aminoacides et les acides aminés contenant des chaînes latérales modifiées ou synthétiques. Les acides aminés seront considérés comme des L‑aminoacides non modifiés sauf s’il est précisé dans leur description qu’ils sont modifiés au sens du paragraphe 29;
3. “vocabulaire contrôlé” désigne la terminologie employée dans la présente norme, qui doit être reprise dans la description des caractéristiques d’une séquence, c’est‑à‑dire dans les annotations de régions ou de sites présentant un intérêt particulier conformément à l’annexe I;
4. “séquence délibérément omise” ou séquence vide désigne un espace réservé qui est destiné à préserver la numérotation des séquences dans le listage afin de garantir la cohérence de cette numérotation avec celle des divulgations jointes à la demande, par exemple lorsqu’une séquence a été supprimée dans la divulgation, pour éviter d’avoir à renuméroter les séquences à la fois dans la divulgation et dans le listage des séquences;
5. “nucléotide” désigne tout nucléotide ou analogue nucléotidique qui peut être représenté à l’aide des symboles indiqués dans l’annexe I (voir section 1, tableau 1). Les nucléotides peuvent notamment contenir une base pyrimidique ou purine modifiée ou synthétique, ou un ribose ou désoxyribose modifié ou synthétique, et peuvent être reliés par une liaison inter‑nucléoside de 3’ à 5’ modifiée ou synthétique, c’est‑à‑dire par toute fraction chimique assurant la même fonction structurelle que la fraction phosphate de l’ADN ou de l’ARN, par exemple la fraction phosphorothioate;
6. “résidu” désigne tout nucléotide ou acide aminé individuel dans une séquence;
7. “numéro d’identification de séquence” désigne un numéro unique (nombre entier) attribué à chaque séquence du listage;
8. “listage des séquences” désigne une partie de la description, dans la demande de brevet déposée ou dans un document déposé après la demande, qui présente la ou les séquences de nucléotides et/ou d’acides aminés divulguées, ainsi que toute autre description complémentaire;
9. “spécialement défini” désigne tout nucléotide différent de ceux qui sont représentés par le symbole “n” et tout acide aminé différent de ceux qui sont représentés par le symbole “X” dans l’annexe I;
10. “inconnu”, pour un nucléotide ou un acide aminé, signifie qu’un seul nucléotide ou acide aminé est présent mais que son identité est inconnue ou non divulguée.

## PORTÉE

1. La présente norme définit les exigences en matière de présentation des listages des séquences de nucléotides et d’acides aminés pour les séquences divulguées dans les demandes de brevet.
2. Un listage des séquences conforme à cette norme (ci‑après “listage des séquences”) contient une partie consacrée aux informations générales et une partie destinée aux données des séquences. Le listage des séquences doit être présenté dans un fichier unique qui doit être au format XML et être conforme à la définition de type de document (DTD) présentée dans l’annexe II. Les informations bibliographiques figurant dans la partie consacrée aux informations générales sont uniquement destinées à associer le listage des séquences à la demande de brevet pour laquelle le listage a été communiqué. La partie consacrée aux données des séquences se compose d’un ou plusieurs éléments de données, chacun d’eux contenant des informations sur une seule séquence. Ces éléments de données des séquences comportent différentes clés de caractérisation et des qualificateurs ultérieurs conformes aux exigences de la Collaboration internationale sur les bases de données de séquences de nucléotides (INSDC) et d’UniProt.
3. Aux fins de la présente norme, une séquence doit être intégrée dans un listage si elle est divulguée dans n’importe quelle partie d’une demande de brevet par l’énumération de ses résidus, et s’il s’agit :
4. d’une séquence non ramifiée ou d’une partie linéaire d’une séquence ramifiée contenant au moins 10 nucléotides définis de manière spécifique, et dont les nucléotides adjacents ont des liaisons de 3’ à 5’ (ou de 5’ à 3’), ou
5. d’une séquence non ramifiée ou d’une partie linéaire d’une séquence ramifiée contenant au moins quatre acides aminés définis de manière spécifique, et dont les acides aminés adjacents ont des liaisons peptidiques.
6. Un listage des séquences ne doit contenir aucune séquence comportant moins de 10 nucléotides définis de manière spécifique ou moins de quatre acides aminés définis de manière spécifique.

## RÉFÉRENCES

1. Les normes et ressources suivantes sont pertinentes à l’égard de la présente norme :

Collaboration internationale sur les   
bases de données de séquences

de nucléotides (INSDC) <http://www.insdc.org/>;

Norme ISO 639-1 – Codes pour   
la représentation des noms de langue Partie 1 : Code Alpha2;

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

Norme du W3C sur le XML 1.0 <http://www.w3.org/>;

Norme [ST.2](http://www.wipo.int/standards/fr/pdf/03-02-01.pdf) de l’OMPI Indication normalisée des dates à l’aide du calendrier grégorien;

Norme [ST.3](http://www.wipo.int/standards/fr/pdf/03-03-01.pdf) de l’OMPI Codes à deux lettres pour la représentation des États, autres entités et organisations intergouvernementales.

Norme [ST.16](http://www.wipo.int/standards/fr/pdf/03-16-01.pdf) de l’OMPI Identification de différents types de documents de brevet;

Norme [ST.25](http://www.wipo.int/standards/fr/pdf/03-25-01.pdf) de l’OMPI Présentation des listages des séquences de nucléotides et d’acides aminés.

## PRÉSENTATION DES SÉQUENCES

1. À chaque séquence doit être attribué un numéro d’identification de séquence distinct. Ces numéros doivent commencer par le chiffre 1 et être incrémentés de manière consécutive par des nombres entiers. Si aucune séquence ne correspond à un numéro d’identification donné, par exemple en cas de séquence délibérément omise, il convient d’insérer la chaîne de caractères “000” à la place de la séquence (voir le paragraphe 58). Le nombre total de séquences doit être indiqué dans le listage des séquences et doit être égal au nombre total de numéros d’identification de séquence, que ces numéros soient suivis d’une séquence ou de la chaîne de caractères “000”.

### *Séquences de nucléotides*

1. Toute séquence de nucléotides doit être représentée par un seul brin de codage, dans le sens 5’‑3’ et de gauche à droite. Les désignations 5’ et 3’ ne doivent pas apparaître dans la séquence. Toute séquence de nucléotides représentée par deux brins de codage et divulguée par énumération des résidus des deux brins doit être présentée sous la forme :
2. d’une seule séquence ou de deux séquences distinctes, chacune disposant de son propre numéro d’identification de séquence, si les deux brins distincts sont entièrement complémentaires l’un de l’autre; ou
3. de deux séquences distinctes, chacune disposant de son propre numéro d’identification de séquence, si les deux brins ne sont pas entièrement complémentaires l’un de l’autre.
4. La numérotation des positions des nucléotides doit commencer par la première base de la séquence, qui portera le numéro 1. Elle doit être continue dans toute la séquence dans le sens 5’‑3’.
5. Cette méthode de numérotation des séquences de nucléotides s’applique aussi aux séquences de nucléotides de configuration circulaire. Dans ce cas, le déposant doit choisir le nucléotide correspondant au premier numéro.
6. Tous les nucléotides d’une séquence doivent être représentés à l’aide des symboles indiqués dans l’annexe I (voir section 1, tableau 1). Seules les lettres minuscules sont autorisées. Tout symbole employé pour représenter un nucléotide ne peut être l’équivalent que d’un seul résidu.
7. Le symbole “t” désigne la thymine dans de l’ADN et l’uracile dans de l’ARN. L’uracile dans de l’ADN ou la thymine dans de l’ARN sont considérés comme des nucléotides modifiés et doivent être accompagnés d’une description supplémentaire au sens du paragraphe 18.
8. Lorsqu’il convient d’employer un symbole ambigu (représentant deux bases possibles ou plus), il faut choisir le symbole le plus restrictif. Si par exemple une base dans une position quelconque pouvait être désignée par “a” ou “g”, il faut employer “r” au lieu de “n”. Le symbole “n” sera considéré comme équivalent à l’un des symboles “a”, “c”, “g” ou “t/u”, sauf s’il est accompagné d’une description supplémentaire au sens des paragraphes 16 et 17 ou 20. Ce symbole “n” ne peut être employé que pour représenter un nucléotide. Il peut représenter un seul nucléotide modifié ou “inconnu” s’il est accompagné d’une description supplémentaire au sens des paragraphes 16 et 17 ou 20.
9. Les nucléotides modifiés doivent être représentés dans la séquence comme les bases non modifiées correspondantes, c’est‑à‑dire par “a”, “c”, “g” ou “t” chaque fois que possible. Tout nucléotide modifié apparaissant dans une séquence et ne pouvant être représenté à l’aide d’un autre symbole indiqué dans l’annexe I (voir section 1, tableau 1), comme par exemple un nucléotide n’existant pas à l’état naturel, doit être représenté par le symbole “n”. Lorsque ce symbole “n” est employé pour représenter un nucléotide modifié, il n’est l’équivalent que d’un seul résidu.
10. Tout nucléotide modifié doit être accompagné d’une description supplémentaire dans le tableau de caractéristiques (voir les paragraphes 59 et suivants) comportant la clé de caractérisation “modified\_base” et le qualificateur obligatoire “mod\_base”. La valeur qualificative ne peut être constituée que d’une seule abréviation issue de l’annexe I (voir section 2, tableau 2). Si cette abréviation est “OTHER”, le nom complet non abrégé de la base modifiée doit être indiqué dans un qualificateur de type “note”. Les abréviations (ou les noms complets) indiquées dans l’annexe I (voir section 2, tableau 2) qui sont mentionnées ci‑dessus ne doivent pas être employées dans la séquence elle‑même.
11. L’uracile dans de l’ADN ou la thymine dans de l’ARN sont considérés comme des nucléotides modifiés et doivent être représentés dans la séquence par un “t” et être accompagnés d’une description supplémentaire dans le tableau de caractéristiques. Cette description doit comporter la clé de caractérisation “modified\_base”, le qualificateur “mod\_base” dont la valeur doit être “OTHER”, et un qualificateur de type “note” dont la valeur doit être respectivement “uracil” ou “thymine”.
12. Les exemples ci‑après illustrent la manière dont des nucléotides modifiés doivent être présentés pour être conformes aux paragraphes 16 et 17 ci‑dessus :

Exemple 1 : Nucléotide modifié représenté par une abréviation indiquée dans l’annexe I (voir section 2, tableau 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>

Exemple 2 : Nucléotide modifié représenté par la valeur “OTHER” indiquée dans l’annexe I (voir section 2, tableau 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. Tout nucléotide “inconnu” doit être représenté par le symbole “n” dans la séquence. Un nucléotide “inconnu” doit en outre être accompagné d’une description supplémentaire dans le tableau de caractéristiques (voir les paragraphes 60 et suivants) comportant la clé de caractérisation “unsure”. Le symbole “n” ne peut être l’équivalent que d’un seul résidu.
2. Toute région contenant un nombre connu de résidus “a”, “c”, “g”, “t” ou “n” auxquels la même description s’applique peut faire l’objet d’une description commune au moyen de la syntaxe “x..y” dans le descripteur d’emplacement de l’élément INSDFeature\_location (voir les paragraphes 65 à 72). On trouvera des détails sur la présentation des variantes de séquence, par exemple des suppressions, des adjonctions ou des remplacements, aux paragraphes 92 à 97.
3. L’exemple ci‑après illustre la présentation d’une région de nucléotides modifiés faisant l’objet de la même description au sens du paragraphe 21 ci‑dessus :

<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>

### *Séquences d’acides aminés*

1. Les acides aminés d’une séquence protéinique ou peptidique doivent être énumérés dans le sens amino‑carboxy et de gauche à droite. Les groupes amino et carboxy ne doivent pas être représentés dans la séquence.
2. La numérotation des positions des acides aminés doit commencer au premier acide aminé de la séquence, numéroté 1, en tenant compte des acides aminés précédant la protéine mature, par exemple les préséquences, les proséquences et les pré‑proséquences ainsi que les séquences signal. Elle doit être continue sur l’ensemble de la séquence dans le sens amino‑carboxy.
3. Tous les acides aminés d’une séquence doivent être représentés à l’aide des symboles indiqués dans l’annexe I (voir section 3, tableau 3). Seules les lettres majuscules sont autorisées. Tout symbole employé pour représenter un acide aminé ne peut être l’équivalent que d’un seul résidu.
4. Lorsqu’il convient d’employer un symbole ambigu (représentant deux acides aminés possibles ou plus), il faut choisir le symbole le plus restrictif. Si par exemple un acide aminé dans une position quelconque pouvait être de l’acide aspartique ou de l’asparagine, il faut employer le symbole “B” au lieu de “X”. Le symbole “X” ne sera pas considéré comme équivalent à l’un des symboles “A”, “R”, “N”, “D”, “C”, “Q”, “E”, “G”, “H”, “I”, “L”, “K”, “M”, “F”, “P”, “O”, “S”, “U”, “T”, “W”, “Y” ou “V”, sauf s’il est accompagné d’une description supplémentaire au sens des paragraphes 28 à 30 ou 31 à 33. Le symbole “X” ne peut être employé que pour représenter un acide aminé. Il peut représenter un seul acide aminé s’il est accompagné d’une description supplémentaire au sens des paragraphes 28 à 30 ou 31 à 33. On trouvera des détails sur la présentation des variantes de séquence, par exemple des suppressions, des adjonctions ou des remplacements, aux paragraphes 92 à 97.
5. Les séquences d’acides aminés séparées par un ou plusieurs espaces blancs ou symboles internes de fin (par exemple “Ter”, l’astérisque ou le point) doivent être présentées comme des séquences distinctes pour chaque séquence qui contient au moins quatre acides aminés définis de manière spécifique et qui est visée par le paragraphe 6. Chacune de ces séquences distinctes doit être présentée, dans le listage des séquences, avec son propre numéro d’identification de séquence et uniquement à l’aide des symboles indiqués dans l’annexe I (voir section 3, tableau 3). Les symboles de fin et les espaces blancs ne doivent pas être employés dans les séquences figurant dans un listage.
6. Les acides aminés modifiés, y compris les D‑aminoacides, doivent être représentés dans la séquence comme les acides aminés non modifiés correspondants chaque fois que possible. Tout acide aminé modifié apparaissant dans une séquence et ne pouvant être représenté à l’aide d’un autre symbole indiqué dans l’annexe I (voir section 3, tableau 3), doit être représenté par le symbole “X”. Ce symbole “X” n’est l’équivalent que d’un seul résidu.
7. Tout acide aminé modifié doit être accompagné d’une description supplémentaire dans le tableau de caractéristiques (voir les paragraphes 60 et suivants). Il convient d’employer la clé de caractérisation “MOD\_RES” et le qualificateur “NOTE” pour les acides aminés modifiés après traduction, et la clé de caractérisation “SITE” ainsi que le qualificateur “NOTE” pour les autres acides aminés modifiés. La valeur du qualificateur “NOTE” doit être soit une abréviation indiquée dans l’annexe I (voir section 4, tableau 4), soit le nom complet non abrégé de l’acide aminé modifié. Les abréviations indiquées dans le tableau 4 précité ou les noms complets non abrégés ne doivent pas être employés dans la séquence elle‑même.
8. Les exemples ci‑après illustrent la manière dont des acides aminés modifiés doivent être présentés pour être conformes au paragraphe 29 ci‑dessus :

Exemple 1 : Acide aminé modifié après traduction

<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>

Exemple 2 : Acide aminé modifié différemment

<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>

Exemple 3 : D‑aminoacide

<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. Tout acide aminé “inconnu” ou “autre” qui n’est pas visé par le paragraphe 28 doit être représenté par le symbole “X” dans la séquence. Le symbole “X” ne peut être l’équivalent que d’un seul résidu.
2. Tout acide aminé “inconnu” désigné par “X” doit être accompagné d’une description supplémentaire dans le tableau de caractéristiques (voir paragraphes 60 et suivants) comportant la clé de caractérisation “UNSURE” et éventuellement le qualificateur “NOTE”. Tout acide aminé “autre” désigné par “X” doit être accompagné d’une description supplémentaire comportant la clé de caractérisation “SITE” ou “MOD\_RES”, selon le cas, ainsi que le qualificateur “NOTE” avec le nom complet non abrégé de l’acide aminé “autre”.
3. Les exemples ci‑après illustrent la manière dont des acides aminés “inconnus” ou “autres” doivent être présentés pour être conformes aux paragraphes 31 et 32 ci‑dessus :

Exemple 1 : Acide aminé “inconnu”

<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>

Exemple 2 : Acide aminé “autre”

<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. Toute région contenant un nombre connu de résidus “X” contigus auxquels la même description s’applique peut faire l’objet d’une description commune au moyen de la syntaxe “x..y” dans le descripteur d’emplacement de l’élément INSDFeature\_location (voir les paragraphes 65 à 71). On trouvera des détails sur la présentation des variantes de séquence, par exemple des suppressions, des adjonctions ou des remplacements, aux paragraphes 92 à 97.

### *Présentation de cas particuliers*

1. Toute séquence divulguée par énumération de ses résidus qui est construite comme une séquence continue et unique d’un ou plusieurs segments non contigus provenant d’une séquence plus grande ou de segments provenant de différentes séquences doit être ajoutée au listage des séquences comme une séquence unique avec un numéro d’identification de séquence unique.
2. Toute séquence divulguée par énumération de ses résidus qui contient des régions de résidus énumérés de manière spécifique et séparés par une ou plusieurs régions de résidus “n” ou “X” contigus (voir respectivement les paragraphes 15 et 26), et pour laquelle le nombre exact de résidus dans chaque région est divulgué, doit être ajoutée au listage des séquences comme une séquence unique avec un numéro d’identification de séquence unique.
3. Toute séquence divulguée par énumération de ses résidus qui contient des régions de résidus énumérés de manière spécifique et séparés par une ou plusieurs brèches composées d’un nombre inconnu ou non divulgué de résidus doit être insérée dans le listage des séquences comme une série de séquences distinctes. Chacune de ces séquences distinctes doit comporter une région de résidus énumérés de manière spécifique et disposer de son propre numéro d’identification de séquence. Le nombre de séquences distinctes doit ainsi être égal au nombre de régions de résidus énumérés de manière spécifique. Les séquences contenant des brèches composées d’un nombre inconnu ou non divulgué de résidus ne doivent pas être ajoutées au listage des séquences en tant que séquences uniques.

## STRUCTURE DU LISTAGE DE SÉQUENCES EN XML

1. En application du paragraphe 5 ci‑dessus, l’instance XML d’un fichier de listage des séquences conforme à la présente norme se compose des éléments suivants :
2. une partie consacrée aux informations générales, qui contient des informations sur la demande de brevet à laquelle se rapporte le listage des séquences; et
3. une partie consacrée aux données des séquences, qui contient un ou plusieurs éléments de données sur les séquences, chacun de ces éléments contenant des informations sur une seule séquence.

On trouvera un exemple de listage de séquences dans l’annexe III.

1. Le listage des séquences doit être présenté au format XML 1.0 en employant la DTD définie dans l’annexe II intitulée “Définition de type de document pour le listage des séquences”.
2. La première ligne de l’instance XML doit contenir la déclaration du format XML :

<?xml version=“1.0” encoding=“UTF‑8”?>.

1. La deuxième ligne de l’instance XML doit contenir une déclaration de type de document (DOCTYPE) :

<!DOCTYPE ST26SequenceListing PUBLIC “‑//WIPO//DTD Sequence Listing 1.0//EN” “ST26SequenceListing\_V1\_0.dtd”>.

1. Le listage des séquences au format électronique doit être entièrement contenu dans un seul fichier. Celui‑ci doit être codé selon la norme Unicode UTF‑8, avec les restrictions suivantes :
2. les informations figurant dans les éléments ApplicantName, InventorName et InventionTitle dans la partie consacrée aux informations générales peuvent comporter n’importe quel caractère Unicode à l’exception des caractères réservés, qui doivent être remplacés selon la méthode décrite au paragraphe 41;
3. les informations figurant dans tous les autres éléments de la partie consacrée aux informations générales et dans tous les éléments de la partie sur les données des séquences
   * doivent être composées de caractères imprimables (y compris le caractère d’espace) indiqués dans le tableau de codes des caractères latins de base de la norme Unicode, à l’exception des caractères réservés, qui doivent être remplacés selon la méthode décrite au paragraphe 41 (c’est‑à‑dire que ces caractères sont limités aux points de codes Unicode 0020, 0021, 0023 jusqu’à 0026, 0028 jusqu’à 003B, 003D, et 003F jusqu’à 007E – voir l’annexe IV), et
   * les seules entités de caractères autorisées sont les entités prédéfinies prévues au paragraphe 41.
4. Dans l’instance XML d’un listage des séquences, les caractères réservés suivants doivent être remplacés par les entités prédéfinies correspondantes lorsqu’ils sont employés pour renseigner la valeur d’un attribut ou le contenu d’un élément :

|  |  |
| --- | --- |
| **Caractère réservé** | **Entités prédéfinies** |
| < | &lt; |
| > | &gt; |
| & | &amp; |
| “ | &quot; |
| ‘ | &apos; |

On trouvera un exemple au paragraphe 72.

1. Tous les éléments obligatoires doivent être renseignés (sauf ceux qui sont indiqués au paragraphe 58 à propos des séquences délibérément omises). Les éléments facultatifs pour lesquels aucun contenu n’est disponible ne doivent pas figurer dans l’instance XML.

### *Élément racine*

1. L’élément racine d’une instance XML au sens de la présente norme est l’élément ST26SequenceListing, dont les attributs sont les suivants :

|  |  |  |
| --- | --- | --- |
| **Attribut** | **Description** | **Obligatoire/Facultatif** |
| dtdVersion | Version de la DTD employée pour créer ce fichier au format “V#\_#”, par exemple “V1\_0”. | Obligatoire |
| fileName | Nom du fichier contenant le listage des séquences. | Facultatif |
| softwareName | Nom du logiciel ayant créé le fichier. | Facultatif |
| softwareVersion | Version du logiciel ayant créé le fichier. | Facultatif |
| productionDate | Date de production du fichier contenant le listage des séquences (format “SSAA‑MM‑JJ”). | Facultatif |

1. L’exemple ci‑après est une illustration de l’élément racine ST26SequenceListing et de ses attributs dans une instance XML conforme au paragraphe 43 ci‑dessus :

<ST26SequenceListing dtdVersion=“V1\_0” fileName=“US11\_405455\_SEQL.xml” softwareName=“SEQL‑software‑name” softwareVersion=“1.0” productionDate=“2006‑05‑10”>

{...}\*

</ST26SequenceListing>

\*{...} représente la partie des informations générales et la partie des données de séquences qui ne figurent pas dans cet exemple.

### *Partie consacrée aux informations générales*

1. Les éléments de la partie consacrée aux informations générales contiennent des informations sur la demande de brevet, comme indiqué ci‑après :

| **Élément** | **Description** | **Obligatoire/**  **Facultatif** |
| --- | --- | --- |
| ApplicationIdentification  L’élément ApplicationIdentification est composé des éléments suivants : | Identification de la demande pour laquelle le listage des séquences est soumis. | Obligatoire lorsqu’un listage des séquences est remis à un moment quelconque après l’attribution d’un numéro de demande. |
| IPOfficeCode | Code ST.3 de l’office de dépôt | Obligatoire |
| ApplicationNumberText | Identification de la demande fournie par l’office de dépôt (ex : PCT/IB2013/099999) | Obligatoire |
| FilingDate | Date de dépôt de la demande de brevet pour laquelle le listage des séquences est remis (au format ST.2 “SSAA‑MM‑JJ”, en désignant l’année civile sur 4 chiffres, le mois civil sur 2 chiffres et le jour du mois civil sur 2 chiffres, p. ex. 2015‑01‑31) | Obligatoire lorsqu’un listage des séquences est remis à un moment quelconque après l’attribution d’une date de dépôt. |
| ApplicantFileReference | Identificateur unique attribué par le demandeur pour désigner une demande particulière, composé de caractères définis au paragraphe 40 b). | Obligatoire lorsqu’un listage des séquences est remis à un moment quelconque avant l’attribution du numéro de demande; facultatif dans les autres cas. |
| EarliestPriorityApplicationIdentification | Identification de la première revendication de priorité dans la demande (contient également les éléments IPOfficeCode, ApplicationNumberText et FilingDate, voir ApplicationIdentification ci‑dessus) | Obligatoire si une priorité est revendiquée. |
| ApplicantName | Nom du premier déposant mentionné, composé de caractères définis au paragraphe 40 a). Cet élément comporte l’attribut obligatoire languageCode conformément au paragraphe 47. | Obligatoire |
| ApplicantNameLatin | Si l’élément ApplicantName comporte des caractères différents de ceux définis au paragraphe 40 b), une traduction ou une translittération du nom du premier déposant mentionné doit être fournie et doit aussi se composer de caractères définis au paragraphe 40 b). | Obligatoire si l’élément ApplicantName contient des caractères non latins. |
| InventorName | Nom du premier inventeur mentionné, composé de caractères définis au paragraphe 40 a). Cet élément comporte l’attribut obligatoire languageCode conformément au paragraphe 47. | Facultatif |
| InventorNameLatin | Si l’élément InventorName comporte des caractères différents de ceux définis au paragraphe 40 b), une traduction ou une translittération du nom du premier inventeur mentionné doit être fournie et doit aussi se composer de caractères définis au paragraphe 40 b). | Facultatif |
| InventionTitle | Titre de l’invention, composé de caractères définis au paragraphe 40 a) dans la langue de dépôt. Une traduction du titre de l’invention dans d’autres langues peut être fournie; elle doit alors se composer de caractères définis au paragraphe 40 a) et apparaître sous des éléments InventionTitle supplémentaires. Cet élément comporte l’attribut obligatoire languageCode défini au paragraphe 48.  Le titre de l’invention doit comporter de préférence deux à sept mots. | Obligatoire dans la langue de dépôt. Facultatif dans d’autres langues. |
| SequenceTotalQuantity | Nombre total de toutes les séquences apparaissant dans le listage, y compris les séquences délibérément omises (également appelées séquences vides) (voir le paragraphe 9). | Obligatoire |

1. Les exemples ci‑après illustrent la manière dont la partie du listage des séquences consacrée aux informations générales doit être présentée pour être conforme au paragraphe 45 ci‑dessus :

Exemple 1 : Listage des séquences déposé avant l’attribution du numéro d’identification et de la date de dépôt de la demande

<?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>

\*{...} représente des informations pertinentes pour chaque séquence qui ne figurent pas dans cet exemple.

Exemple 2 : Listage des séquences déposé après l’attribution du numéro d’identification et de la date de dépôt de la demande

<?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>

\*{...} représente des informations pertinentes pour chaque séquence qui ne figurent pas dans cet exemple.

1. Le nom du déposant et, à titre facultatif, le nom de l’inventeur doivent être indiqués respectivement dans les éléments ApplicantName et InventorName car ils sont généralement mentionnés dans la langue de dépôt de la demande. Le code de langue adéquat (voir le paragraphe 8 b)) doit être indiqué dans l’attribut languageCode pour chaque élément. Si le nom du déposant contient des caractères différents de l’alphabet latin tel que défini au paragraphe 40 b), une translittération ou une traduction de ce nom doit aussi être fournie en caractères latins dans l’élément ApplicantNameLatin. Si le nom de l’inventeur contient des caractères différents de l’alphabet latin, une translittération ou une traduction de ce nom doit aussi être fournie en caractères latins dans l’élément InventorNameLatin.
2. Le titre de l’invention doit être indiqué dans l’élément InventionTitle dans la langue de dépôt et peut aussi figurer dans d’autres langues en ajoutant d’autres éléments InventionTitle (voir le tableau du paragraphe 45). Le code de langue adéquat (voir le paragraphe 8 b)) doit être indiqué dans l’attribut languageCode de l’élément.
3. L’exemple ci‑après illustre la manière dont les noms et le titre de l’invention doivent être présentés pour être conformes aux paragraphes 47 et 48 ci‑dessus :

Exemple : Le nom du déposant et celui de l’inventeur sont présentés en caractères japonais et latins et le titre de l’invention est présenté en japonais, en anglais et en français

<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>

### *Partie consacrée aux données sur les séquences*

1. La partie consacrée aux données sur les séquences doit être composée d’un ou plusieurs éléments SequenceData, chacun d’eux contenant des informations sur une séquence.
2. Chaque élément SequenceData doit avoir un attribut obligatoire sequenceIDNumber contenant le numéro d’identification de la séquence (voir le paragraphe 9), par exemple :

<SequenceData sequenceIDNumber=“1”>

1. L’élément SequenceData doit contenir un élément subordonné INSDSeq qui se compose d’autres éléments subordonnés, de la manière suivante :

| **Élément** | **Description** | **Obligatoire/Non indiqué** | |
| --- | --- | --- | --- |
| **Séquences** | **Séquences délibérément omises** |
| INSDSeq\_length | Longueur de la séquence | Obligatoire | Obligatoire, aucune valeur indiquée |
| INSDSeq\_moltype | Type de molécule | Obligatoire | Obligatoire, aucune valeur indiquée |
| INSDSeq\_division | Indication du fait qu’une séquence est liée à une demande de brevet | Obligatoire avec la valeur “PAT” | Obligatoire, aucune valeur indiquée |
| INSDSeq\_feature‑table | Liste d’annotations de la séquence | Obligatoire | Ne doit PAS être indiqué |
| INSDSeq\_sequence | Séquence | Obligatoire | Obligatoire, indiquer la valeur “000” |

1. L’élément INSDSeq\_length doit divulguer le nombre de nucléotides ou d’acides aminés de la séquence figurant dans l’élément INSDSeq\_sequence, par exemple :

<INSDSeq\_length>8</INSDSeq\_length>

1. L’élément INSDSeq\_moltype doit divulguer le type de la molécule présentée. Pour les séquences de nucléotides, le type de molécule doit être ADN ou ARN. Pour les séquences protéiniques ou peptidiques, le type de molécule doit être AA. (Cet élément est distinct des qualificateurs “mol\_type” et “MOL\_TYPE” mentionnés aux paragraphes 55 et 85.) Par exemple :

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

1. Si une séquence de nucléotides contient à la fois des fragments d’ADN et d’ARN, l’élément INSDSeq\_moltype doit prendre la valeur “DNA”. La molécule combinée d’ADN/ARN doit en outre être décrite dans le tableau de caractéristiques à l’aide de la clé de caractérisation “source”, du qualificateur obligatoire “organism”, qui prend la valeur “synthetic construct”, et du qualificateur obligatoire “mol\_type”, qui prend la valeur “other DNA”. Chaque fragment d’ADN et d’ARN de la molécule combinée d’ADN/ARN doit en outre être décrit par la clé de caractérisation “misc\_feature” et par le qualificateur “note”, ce dernier indiquant s’il s’agit d’un fragment d’ADN ou d’ARN.
2. L’exemple ci‑après illustre la description d’une séquence de nucléotides contenant à la fois des fragments d’ADN et d’ARN comme le prévoit le paragraphe 55 ci‑dessus :

<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. L’élément INSDSeq\_sequence doit divulguer la séquence. Les résidus de la séquence doivent être présentés de manière contiguë à l’aide des symboles adéquats indiqués dans l’annexe I (voir section 1, tableau 1 et section 3, tableau 3). La séquence ne doit pas contenir de chiffres, de signes de ponctuation ou d’espaces blancs.
2. Une séquence délibérément omise doit être présentée de la manière suivante :
3. l’élément SequenceData et son attribut sequenceIDNumber, qui prend pour valeur le numéro d’identification de la séquence omise;
4. les éléments INSDSeq\_length, INSDSeq\_moltype, INSDSeq\_division, qui sont présents mais ne contiennent aucune valeur;
5. l’élément INSDSeq\_feature‑table ne doit pas être indiqué; et
6. l’élément INSDSeq\_sequence, qui prend la valeur “000”.
7. L’exemple ci‑après illustre la manière dont une séquence délibérément omise doit être présentée pour être conforme au paragraphe 58 ci‑dessus :

<SequenceData sequenceIDNumber=“3”>

<INSDSeq>

<INSDSeq\_length/>

<INSDSeq\_moltype/>

<INSDSeq\_division/>

<INSDSeq\_sequence>000</INSDSeq\_sequence>

</INSDSeq>

</SequenceData>

### *Tableau de caractéristiques*

1. Le tableau de caractéristiques contient des informations sur l’emplacement et les rôles des différentes régions d’une séquence particulière. Il est obligatoire de fournir un tableau de caractéristiques pour chaque séquence, sauf s’il s’agit d’une séquence délibérément omise; dans ce cas, ce tableau ne doit pas apparaître. Le tableau de caractéristiques figure dans l’élément INSDSeq\_feature‑table, qui se compose d’un ou plusieurs éléments INSDFeature.
2. Chaque élément INSDFeature contient la description d’une caractéristique et se compose d’éléments subordonnés de la manière suivante :

|  |  |  |
| --- | --- | --- |
| **Élément** | **Description** | **Obligatoire/Facultatif** |
| INSDFeature\_key | Mot ou abréviation indiquant une caractéristique | Obligatoire |
| INSDFeature\_location | Région de la séquence présentée correspondant à la caractéristique | Obligatoire |
| INSDFeature\_quals | Qualificateur contenant des informations complémentaires sur une caractéristique | Obligatoire si la clé de caractérisation nécessite un ou plusieurs qualificateurs, p. ex. “source”. Facultatif dans les autres cas. |

### *Clés de caractérisation*

1. L’annexe I contient une liste complète des clés de caractérisation qui peuvent être employées dans le cadre de la présente norme, ainsi qu’une liste complète des qualificateurs associés à ces clés, dans laquelle il est précisé si les qualificateurs sont obligatoires ou facultatifs. La section 5 de l’annexe I contient la liste complète des clés de caractérisation destinées aux séquences de nucléotides, et la section 7 contient la liste complète des clés de caractérisation destinées aux séquences d’acides aminés.

### *Clés de caractérisation obligatoires*

1. La clé de caractérisation “source” est obligatoire pour toutes les séquences de nucléotides et la clé de caractérisation “SOURCE” est obligatoire pour toutes les séquences d’acides aminés, sauf s’il s’agit d’une séquence délibérément omise. Chaque séquence doit comporter une clé “source” ou “SOURCE” unique couvrant la séquence tout entière. Si une séquence provient de plusieurs sources, celles‑ci doivent en outre être décrites dans le tableau de caractéristiques à l’aide de la clé de caractérisation “misc\_feature” et du qualificateur “note” pour les séquences de nucléotides, et de la clé de caractérisation “REGION” et du qualificateur “NOTE” pour les séquences d’acides aminés.
2. Certaines clés de caractérisation nécessitent l’emploi d’une clé de caractérisation supplémentaire, appelée “Parent Key”. Ainsi, la clé de caractérisation “C\_region” doit être accompagnée de la clé “CDS” (voir annexe I, section 5).

### *Emplacement de la caractéristique*

1. L’élément obligatoire INSDFeature\_location doit contenir au moins un descripteur d’emplacement qui définit un site ou une région correspondant à une caractéristique de la séquence dans l’élément INSDSeq\_sequence. Il peut contenir plusieurs descripteurs d’emplacement (voir les paragraphes 68 à 71).
2. Le descripteur d’emplacement peut être un numéro de résidu unique, un site entre deux numéros de résidus adjacents, une région délimitant une série de numéros de résidus contigus, ou un site ou une région qui s’étend au‑delà du résidu ou de la série de résidus particuliers. On peut employer plusieurs descripteurs d’emplacement en conjonction avec un opérateur d’emplacement quand une caractéristique correspond à des sites ou des régions discontinus de la séquence (voir les paragraphes 68 à 71). Le descripteur d’emplacement ne doit pas comporter de numéros de résidus en dehors de la série indiquée pour la séquence dans l’élément INSDSeq\_sequence.
3. La syntaxe de chaque type de descripteurs d’emplacement est indiquée dans le tableau ci‑dessous, où x et y sont des numéros de résidus indiqués en nombres entiers non négatifs et inférieurs ou égaux à la longueur de la séquence dans l’élément INSDSeq\_sequence, et où x est inférieur à y.

| **Type de descripteurs d’emplacement** | **Syntaxe** | **Description** |
| --- | --- | --- |
| Numéro de résidu unique | x | Désigne un résidu unique dans la séquence présentée. |
| Numéros de résidus délimitant un ensemble dans la séquence | x..y | Désigne une série continue de résidus délimitée par un résidu de début et un résidu de fin, ces deux résidus étant inclus dans la série. |
| Résidus situés avant le premier ou après le dernier numéro de résidu indiqué | <x  >x  <x..y  x..>y | Désigne une région qui comprend un résidu ou une série de résidus indiqués et qui s’étend au‑delà d’un résidu indiqué. Les symboles “<” et “>” peuvent être employés à l’égard d’un résidu unique ou des numéros du résidu de début et de fin d’une série de résidus pour signaler qu’une caractéristique s’étend au‑delà du numéro de résidu indiqué. |
| Site s’étendant entre deux numéros de résidus adjacents | x^y | Désigne un site entre deux résidus adjacents, par exemple le site d’un clivage endonucléolytique. Les numéros de position des résidus adjacents sont séparés par un caret (^). Les formats autorisés pour ce descripteur sont x^x+1 (par exemple 55^56), ou pour les nucléotides circulaires, x^1, où “x” est la longueur totale de la molécule, c’est‑à‑dire 1000^1 pour une molécule circulaire de longueur 1000. |

1. Un opérateur d’emplacement est le préfixe d’un descripteur ou d’une combinaison de descripteurs d’emplacement correspondant à une caractéristique unique mais discontinue. Il indique l’emplacement correspondant à la caractéristique dans la séquence présentée, ou comment la caractéristique est construite. Une liste d’opérateurs d’emplacement est fournie ci‑après avec leur définition.
2. Opérateur d’emplacement pour des nucléotides et des acides aminés :

|  |  |
| --- | --- |
| **Syntaxe de l’emplacement** | **Description de l’emplacement** |
| join(location,location, ... location) | Les emplacements indiqués sont joints (placés bout à bout) pour former une seule séquence contiguë. |
| order(location,location, ... location) | Les éléments se trouvent dans l’ordre indiqué mais aucune information ne permet de déterminer s’il est raisonnable de les joindre. |

1. Opérateur d’emplacement réservé aux nucléotides :

|  |  |
| --- | --- |
| **Syntaxe de l’emplacement** | **Description de l’emplacement** |
| complement(location) | Indique que la caractéristique se trouve sur le brin complémentaire à la série de la séquence indiquée par le descripteur d’emplacement, lorsque la séquence est lue dans le sens 5’‑3’. |

1. Les opérateurs d’emplacement assurant un rôle de jonction ou d’ordonnancement nécessitent au moins deux descripteurs d’emplacement séparés par des virgules. Les descripteurs d’emplacement concernant des sites situés entre deux résidus adjacents, c’est‑à‑dire x^y, ne peuvent être employés dans un emplacement de jonction ou d’ordonnancement. L’emploi de l’opérateur d’emplacement de jonction implique que les résidus désignés par les descripteurs d’emplacement sont physiquement mis en contact par des processus biologiques (par exemple, les exons qui contribuent à la caractéristique d’une région jouant un rôle de codage).
2. L’opérateur d’emplacement “complement” ne peut être employé que pour des nucléotides. Il peut être employé en combinaison soit avec “join” soit avec “order” dans le même emplacement. Les combinaisons de “join” et “order” ne sont pas autorisées dans un même emplacement.
3. Les exemples ci‑après illustrent des emplacements de caractéristiques au sens des paragraphes 65 à 70 ci‑dessus :
4. emplacements pour des nucléotides et des acides aminés :

| **Exemple d’emplacement** | **Description** |
| --- | --- |
| 467 | Désigne le résidu 467 de la séquence. |
| 123^124 | Désigne un site entre les résidus 123 et 124. |
| 340..565 | Désigne une série continue de résidus dont les bornes sont le 340 et le 565, ces bornes étant incluses dans la série. |
| <1 | Désigne un emplacement de caractéristique situé avant le premier résidu. |
| <345..500 | Indique que le point exact de la borne inférieure d’une caractéristique est inconnu. L’emplacement commence à un résidu situé quelque part avant le 345 et continue jusqu’au résidu 500 inclus. |
| <1..888 | Indique que la caractéristique commence avant le premier résidu de la séquence et continue jusqu’au résidu 888 inclus. |
| 1..>888 | Indique que la caractéristique commence au premier résidu de la séquence et continue au‑delà du résidu 888. |
| join(12..78,134..202) | Indique que les régions 12 à 78 et 134 à 202 devraient être jointes pour constituer une séquence contiguë. |

1. emplacements réservés aux nucléotides :

|  |  |
| --- | --- |
| **Exemple d’emplacement** | **Description** |
| complement(34..126) | Commence à la base complémentaire à la base 126 et finit à la base complémentaire à la base 34 (la caractéristique est située sur le brin complémentaire au brin présenté). |
| complement(join(2691..4571, 4918..5163)) | Joint les bases 2691 à 4571 et 4918 à 5163, puis complète les segments joints (la caractéristique est située sur le brin complémentaire au brin présenté). |
| join(complement(4918..5163), complement(2691..4571)) | Complète les régions 4918 à 5163 et 2691 à 4571, puis joint les segments complétés (la caractéristique est située sur le brin complémentaire au brin présenté). |

1. Dans une instance XML d’un listage des séquences, les caractères “<” et “>” d’un descripteur d’emplacement doivent être remplacés par les entités prédéfinies adéquates (voir le paragraphe 41), par exemple :

Feature location “<1” :

<INSDFeature\_location>&lt;1</INSDFeature\_location>

Feature location “1..>888” :

<INSDFeature\_location>1..&gt;888</INSDFeature\_location>

### *Qualificateurs de caractéristiques*

1. Les qualificateurs permettent de fournir des informations sur les caractéristiques pour compléter les informations figurant dans la clé de caractérisation et l’emplacement de la caractéristique. La valeur des qualificateurs peut prendre trois types de formats selon le type d’informations fournies :
2. du texte libre (voir les paragraphes 86 et 87);
3. un vocabulaire contrôlé ou l’énumération de valeurs (p. ex. un nombre ou une date); et
4. des séquences.
5. La section 6 de l’annexe I contient une liste complète des qualificateurs et la définition du format de leurs valeurs, le cas échéant, pour la clé de caractérisation de chaque nucléotide, et la section 8 contient la liste complète des qualificateurs pour la clé de caractérisation de chaque acide aminé.
6. Toute séquence prévue au paragraphe 6 qui est indiquée à titre de valeur d’un qualificateur doit être présentée de manière distincte dans le listage des séquences avec son propre numéro d’identification de séquence.

### *Qualificateurs de caractéristiques obligatoires*

1. Une clé de caractérisation obligatoire, en l’occurrence “source” pour les séquences de nucléotides et “SOURCE” pour les séquences d’acides aminés, doit être accompagnée de deux qualificateurs obligatoires, “organism” et “mol\_type” pour les séquences de nucléotides et “ORGANISM” et “MOL\_TYPE” pour les séquences d’acides aminés. Certaines clés de caractérisation facultatives nécessitent aussi des qualificateurs obligatoires.

### *Éléments des qualificateurs*

1. L’élément INSDFeature\_quals se compose d’un ou plusieurs éléments INSDQualifier. Chaque élément INSDQualifier représente un seul qualificateur et se compose de deux éléments subordonnés, de la manière suivante :

|  |  |  |
| --- | --- | --- |
| **Élément** | **Description** | **Obligatoire/Facultatif** |
| INSDQualifier\_name | Nom du qualificateur (voir annexe I, sections 6 et 8) | Obligatoire |
| INSDQualifier\_value | Valeur du qualificateur, le cas échéant, au format indiqué (voir annexe I, sections 6 et 8) | Obligatoire si indiqué (voir annexe I, sections 6 et 8) |

1. Le qualificateur d’organisme, c’est‑à‑dire l’élément “organism” pour les séquences de nucléotides (voir annexe I, section 6) et “ORGANISM” pour les séquences d’acides aminés (voir annexe I, section 8) doit divulguer la source, c’est‑à‑dire l’organisme ou l’origine unique de la séquence qui est présentée. Les indications d’organisme doivent être choisies parmi les éléments d’une taxonomie.
2. Si la séquence existe à l’état naturel et qu’il existe une désignation de genre et d’espèce en latin pour l’organisme source, le qualificateur doit prendre cette désignation pour valeur. Il est possible d’indiquer le nom commun anglais le plus courant à l’aide du qualificateur “note” pour les séquences de nucléotides et “NOTE” pour les séquences d’acides aminés, mais ce nom ne doit pas être employé comme valeur du qualificateur d’organisme.
3. Les exemples suivants illustrent la source de séquences présentées conformément aux paragraphes 78 et 79 ci‑dessus :

Exemple 1 : Source d’une séquence de nucléotides

<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>

Exemple 2 : Source d’une séquence de protéines

<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 séquence existe à l’état naturel et qu’il existe une désignation de genre en latin pour l’organisme source, mais que l’espèce n’est pas indiquée ou connue, le qualificateur d’organisme doit prendre pour valeur le genre en latin suivi de “sp.”, par exemple :

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

<INSDQualifier\_value>Bacillus sp.</INSDQualifier\_value>

1. Si la source de la séquence est naturelle, mais que la désignation latine de genre et d’espèce de l’organisme est inconnue, le qualificateur d’organisme doit prendre pour valeur l’indication “unidentified” et être accompagné de toute information taxonomique connue dans le qualificateur “note” pour les séquences de nucléotides et “NOTE” pour les séquences d’acides aminés, par exemple :

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

<INSDQualifier\_value>unidentified</INSDQualifier\_value>

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

<INSDQualifier\_value>bacterium B8</INSDQualifier\_value>

1. Si la séquence existe à l’état naturel et que l’organisme source n’a pas de désignation de genre et d’espèce en latin (comme par exemple un virus), le qualificateur d’organisme peut prendre pour valeur un autre nom scientifique acceptable (ex : “adénovirus canin type 2”), par exemple :

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

<INSDQualifier\_value>Canine adenovirus type 2</INSDQualifier\_value>

1. Si la séquence n’existe pas à l’état naturel, le qualificateur d’organisme doit prendre pour valeur “synthetic construct”. On peut ajouter d’autres informations sur la manière dont la séquence a été créée à l’aide du qualificateur “note” pour les séquences de nucléotides et “NOTE” pour les séquences d’acides aminés, par exemple :

<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. Le qualificateur “mol\_type” pour les séquences de nucléotides (voir annexe I, section 6) et “MOL\_TYPE” pour les séquences d’acides aminés (voir annexe I, section 8) doit divulguer le type de molécule représenté dans la séquence. Ces qualificateurs sont distincts de l’élément INSDSeq\_moltype examiné au paragraphe 54 :
2. pour des séquences de nucléotides, la valeur du qualificateur “mol\_type” doit être l’un des éléments suivants : “genomic DNA”, “genomic RNA”, “mRNA”, “tRNA”, “rRNA”, “other RNA”, “other DNA”, “transcribed RNA”, “viral cRNA”, “unassigned DNA” ou “unassigned RNA”. Si la séquence n’existe pas à l’état naturel, c’est‑à‑dire si la valeur du qualificateur “organism” est “synthetic construct”, la valeur du qualificateur “mol\_type” doit être soit “other RNA” soit “other DNA”;
3. pour des séquences d’acides aminés, la valeur du qualificateur “MOL\_TYPE” est “protein”.

### *Texte libre*

1. Le texte libre est un format de valeur autorisé pour certains qualificateurs (comme indiqué à l’annexe I). Il s’agit d’un texte descriptif qui se présente sous forme de segments de phrases, de préférence en anglais.
2. L’emploi du texte libre doit être limité à un petit nombre de termes brefs indispensables à la compréhension d’une caractéristique de la séquence. Pour chaque qualificateur, le texte libre ne peut compter plus de 1000 caractères.

### *Séquences de codage*

1. La clé de caractérisation “CDS” peut servir à désigner des séquences de codage, c’est‑à‑dire des séquences de nucléotides correspondant à la séquence d’acides aminés dans une protéine et au codon d’arrêt. L’élément INSDFeature\_location doit indiquer l’emplacement de la caractéristique “CDS”, y compris le codon d’arrêt.
2. Les qualificateurs “transl\_table” et “translation” peuvent être employés en association avec la clé de caractérisation “CDS” (voir annexe I). Si le qualificateur “transl\_table” n’est pas employé, on présume que c’est le tableau de codes normalisés qui est appliqué (voir annexe I, section 9, tableau 5).
3. Toute séquence protéinique codée selon la séquence de codage et divulguée dans un qualificateur de type “translation” visé par le paragraphe 6 doit avoir son propre numéro d’identification de séquence et doit être présentée dans le listage des séquences. Le numéro d’identification de séquence attribué à la séquence protéinique doit figurer dans la valeur du qualificateur “protein\_id” associé à la clé de caractérisation “CDS”. Le qualificateur “ORGANISM” associé à la clé de caractérisation “SOURCE” de la séquence protéinique doit être identique à celui de sa séquence de codage, par exemple :

<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. Toute séquence primaire et toute variante de cette séquence, chacune d’elles étant divulguée par énumération de ses résidus et visée par le paragraphe 6, doit être présentée dans le listage des séquences avec son propre numéro d’identification de séquence.
2. Toute séquence de variante, divulguée uniquement par référence à un ou plusieurs suppressions, adjonctions ou remplacements effectués dans une séquence primaire figurant dans le listage des séquences, peut être présentée dans le listage des séquences. Si tel est le cas, cette séquence de variante :
3. peut être présentée par annotation de la séquence primaire, si elle comporte une ou plusieurs variations à un seul emplacement ou à plusieurs emplacements distincts et que les occurrences de ces variations sont indépendantes;
4. devrait être présentée en tant que séquence distincte avec son propre numéro d’identification de séquence, si elle comporte des variations à plusieurs emplacements distincts et que les occurrences de ces variations sont interdépendantes; et
5. doit être présentée en tant que séquence distincte avec son propre numéro d’identification de séquence, si elle comporte une séquence qui a été ajoutée ou remplacée et qui contient plus de 1000 résidus (voir le paragraphe 87).
6. Le tableau ci‑dessous indique le bon usage des clés de caractérisation et des qualificateurs pour des variantes d’acides nucléiques et d’acides aminés :

| **Type de séquence** | **Clé de caractérisation** | **Qualificateur** | **Usage** |
| --- | --- | --- | --- |
| Acide nucléique | variation | replace | Mutations et polymorphismes existant à l’état naturel, p. ex. des allèles ou des polymorphismes de longueur des fragments de restriction |
| Acide nucléique | misc\_difference | replace | La variabilité a été créée artificiellement, p. ex. par une manipulation génétique ou une synthèse chimique |
| Acide aminé | VAR\_SEQ | NOTE | La variante a été produite par un épissage alternatif, l’usage de promoteurs alternatifs, une initiation alternative et un déphasage ribosomique |
| Acide aminé | VARIANT | NOTE | Tout type de variante pour laquelle VAR\_SEQ n’est pas applicable |

1. L’annotation d’une séquence primaire effectuée pour une variante particulière doit comporter une clé de caractérisation et un qualificateur, conformément au tableau ci‑dessus, et indiquer l’emplacement de la caractéristique. Toute suppression doit être représentée par une valeur de qualificateur vide. Tout résidu ajouté ou remplacé doit être indiqué dans le qualificateur “replace” ou “NOTE”. La valeur du qualificateur “replace” ou “NOTE” est un texte libre qui ne doit pas dépasser 1000 caractères, conformément au paragraphe 87. Pour les séquences visées par le paragraphe 6 qui sont présentées à titre d’adjonction ou de remplacement de la valeur d’un qualificateur, se reporter au paragraphe 97. La valeur du qualificateur peut comporter un listage des résidus alternatifs pouvant être ajoutés ou remplacés.
2. Les symboles indiqués dans l’annexe I (voir respectivement les sections 1 à 4, tableaux 1 à 4) peuvent être employés pour représenter des résidus de variantes, selon les besoins. Si un résidu de variante est un résidu modifié qui ne figure pas dans les tableaux 2 ou 4 de l’annexe I, le nom complet non abrégé du résidu modifié doit être indiqué dans la valeur du qualificateur.
3. Les exemples ci‑après illustrent la manière de présenter des variantes pour qu’elles soient conformes aux paragraphes 92 à 95 ci‑dessus :

Exemple 1 : Clé de caractérisation “variation” pour un remplacement dans une séquence de nucléotides.

Une cytosine remplace le nucléotide défini à la position 413 de la séquence.

<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>

Exemple 2 : Clé de caractérisation “misc\_difference” pour une suppression dans une séquence de nucléotides.

Le nucléotide à la position 413 de la séquence est supprimé.

<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>

Exemple 3 : Clé de caractérisation “misc\_difference” pour une adjonction dans une séquence de nucléotides.

La séquence “atgccaaatat” est ajoutée entre les positions 100 et 101 de la séquence primaire.

<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>

Exemple 4 : Clé de caractérisation “VARIANT” pour un remplacement dans une séquence d’acides aminés.

L’acide aminé indiqué à la position 100 de la séquence peut être remplacé par I, A, F, Y, aIle, MeIle, ou 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>

Exemple 5 : Clé de caractérisation “VARIANT” pour un remplacement dans une séquence d’acides aminés.

L’acide aminé indiqué à la position 100 de la séquence peut être remplacé par tout acide aminé sauf Lys, Arg ou 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. Toute séquence visée par le paragraphe 6 qui est présentée à titre d’adjonction ou de remplacement dans la valeur d’un qualificateur pour une annotation de séquence primaire doit aussi être présentée dans le listage des séquences avec son propre numéro d’identification de séquence.

[L’annexe I à la norme ST.26 suit]