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 1.7

Projet présenté pour approbation au Comité des normes de l'OMPI (CWS) à sa onzième session le8 décembre 2023

Note éditoriale du Bureau International

À sa onzième session, le Comité des normes de l'OMPI a décidé que la version 1.7 de la norme ST.26 entrerait en vigueur le 1er juillet 2023. Version 1.6 de la norme ST.26 de l'OMPI doit continuer d'être appliquée jusqu'à l'entrée en vigueur de la nouvelle version.

TABLE DES MATIÈRES

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 1.7

Projet présenté pour approbation au Comité des normes de l'OMPI (CWS) à sa onzième session le8 décembre 2023

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 la façon dont ces divulgations doivent être représentées 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 intellectuelle 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 :

a) 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;

b) 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;

c) de faciliter la recherche de données sur ces séquences; et

d) 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

3. Aux fins de la présente norme, l'expression :

a) "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 acides aminés D et les acides aminés contenant des chaînes latérales modifiées ou synthétiques. Les acides aminés seront considérés comme des acides aminés L non modifiés sauf s'il est précisé dans leur description dans le tableau de caractéristiques qu'ils sont modifiés au sens du paragraphe 30. Aux fins de la présente norme, un résidu d'acide nucléique peptidique (ANP) est considéré non pas comme un acide aminé, mais comme un nucléotide conformément à ce qui est indiqué au paragraphe 3.g)i)2);

b) "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;

c) "énumération de ses résidus" désigne la divulgation d'une séquence dans une demande de brevet sous forme de listage, dans un ordre donné, de chacun des résidus de la séquence, étant entendu que :

i) le résidu est représenté par un nom, une abréviation, un symbole ou une structure (p. ex. HHHHHHQ ou HisHisHisHisHisHisGln); ou

ii) les résidus multiples sont représentés par une formule topologique (p. ex. His_6Gln);

d) "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;

e) "acide aminé modifié" désigne tout acide aminé tel que décrit au paragraphe 3.a) différent de L-alanine, L-arginine, L-asparagine, L-aspartate, L-cystéine, L-glutamine, L-glutamate, L-glycine, L-histidine, L-isoleucine, L-leucine, L-lysine, L-méthionine, L-phénylalanine, L-proline, L-pyrrolysine, L-sérine, L-sélénocystéine, L-thréonine, L-tryptophane, L-tyrosine ou L-valine;

f) "nucléotide modifié" désigne tout nucléotide tel que décrit au paragraphe 3.g) différent de la désoxyadénosine 5'-monophosphate, de la désoxyguanosine 5'-monophosphate, de la désoxycytidine 5'-monophosphate, de la désoxythymidine 5'-monophosphate, de l'adénosine 5'-monophosphate, de la guanosine 5'-monophosphate, de la cytidine 5'-monophosphate ou de l'uridine 5'-monophosphate;

g) "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), le nucléotide ou analogue nucléotidique comprenant :

i) une fraction squelette sélectionnée parmi

1) un 2' désoxyribose 5' monophosphate (la fraction squelette d'un désoxyribonucléotide) ou un ribose 5' monophosphate (la fraction squelette d'un ribonucléotide); ou

2) un analogue du 2' désoxyribose 5' monophosphate ou du ribose 5' monophosphate, qui lorsqu'il constitue le squelette d'un analogue d'acide nucléique, forme une disposition de bases azotées reproduisant celle des acides nucléiques contenant un squelette 2' désoxyribose 5' monophosphate ou ribose 5' monophosphate, l'analogue d'acide nucléique étant capable de former une paire de base avec à un acide nucléique complémentaire; on peut citer comme exemples de fractions squelettes les acides aminés dans les acides nucléiques peptidiques, les molécules de glycol dans les acides nucléiques à glycol, les molécules de sucre thréofuranosyl dans les acides nucléiques à thréose, les cycles morpholiniques et les groupes phosphorodiamidate dans les morpholinos, et les molécules cyclohexényle dans les acides nucléiques à cyclohexényle;

et

ii) le squelette étant

1) relié à une base azotée, y compris une base azotée pyrimidique ou purine modifiée ou synthétique; ou

2) dépourvu d'une base azotée pyrimidique ou purine lorsque le nucléotide fait partie d'une séquence nucléotidique, soit un "site AP" ou "site abasique";

h) "résidu" désigne tout nucléotide ou acide aminé individuel ou leurs analogues respectifs dans une séquence;

i) "numéro d'identification de séquence" désigne un numéro unique (nombre entier) attribué à chaque séquence du listage;

j) "listage des séquences" désigne une partie de la description, dans la demande de brevet déposée ou dans un document déposé après la demande, qui comprend la ou les séquences de nucléotides et/ou d'acides aminés divulguées, ainsi que toute autre description complémentaire, tel que prescrit par la présente norme;

k) "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 (voir section 1, tableau 1, et section 3, tableau 3, respectivement);

l) "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.

m) "séquence variante" désigne une séquence de nucléotides ou d'acides aminés qui présente une ou plusieurs différences par rapport à une séquence primaire. Ces différences peuvent être des résidus alternatifs (voir les paragraphes 15 et 27), des résidus modifiés (voir les paragraphes 3.g), 3.h), 16 et 29), des suppressions, des adjonctions ou des remplacements. Voir les paragraphes 93 à 95.

n) "texte libre" désigne un format de valeur autorisé pour certains qualificateurs. Il s'agit d'un texte descriptif qui se présente sous forme de segments de phrases ou tout autre format précisé (comme indiqué à l'annexe I). Voir le paragraphe 85.

o) "texte libre dépendant de la langue" désigne une valeur de texte libre de certains qualificateurs qui est dépendante de la langue et peut nécessiter une traduction aux fins des procédures internationales, nationales ou régionales. Voir le paragraphe 87.

4. Aux fins de la présente norme,

a) le terme "peut" indique qu'une démarche est facultative ou autorisée, mais pas obligatoire;

b) le terme "doit" indique qu'une démarche est obligatoire selon la présente norme et que le non-respect de celle-ci peut entraîner la non-conformité de la demande;

c) l'expression "ne doit pas" indique une interdiction au sens de la présente norme;

d) le terme "devrait" indique qu'une démarche est fortement conseillée, mais pas obligatoire.

e) l'expression "ne devrait pas" indique qu'une démarche est fortement déconseillée, mais pas interdite.

PORTÉE

5. 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.

6. 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.

7. 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 peut être représentée sous la forme :

a) d'une séquence non ramifiée ou d'une région linéaire d'une séquence ramifiée contenant au moins 10 nucléotides définis de manière spécifique, et dont les nucléotides adjacents sont reliés par :

i) une liaison phosphodiester de 3' à 5' (ou 5' à 3'); ou

ii) toute liaison chimique résultant en une disposition de bases azotées adjacentes qui reproduit la disposition des bases azotées des acides nucléiques existant à l'état naturel; ou

b) d'une séquence non ramifiée ou d'une région linéaire d'une séquence ramifiée contenant au moins quatre acides aminés définis de manière spécifique, et dont les acides aminés forment un squelette peptidique, c'est-à-dire que les acides aminés adjacents ont des liaisons peptidiques.

8. Un listage des séquences ne doit contenir, en tant que séquence disposant de son propre numéro d'identification de séquence, aucune séquence comportant moins de 10 nucléotides définis de manière spécifique ou moins de quatre acides aminés définis de manière spécifique.

RÉFÉRENCES

9. 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 internationale ISO 639−1 :2002	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 de l'OMPI	Indication normalisée des dates à l'aide du calendrier grégorien;
Norme ST.3 de l'OMPI	Norme recommandée concernant les codes à deux lettres pour la représentation des États, autres entités et organisations intergouvernementales.
Norme ST.16 de l'OMPI	Code normalisé recommandé pour l'identification de différents types de documents de brevet;
Norme ST.25 de l'OMPI	Norme relative à la présentation du listage des séquences de nucléotides et d'acides aminés dans les demandes de brevet.

REPRÉSENTATION DES SÉQUENCES

10. À chaque séquence visée par le paragraphe 7 doit être attribué un numéro d'identification de séquence distinct, y compris en ce qui concerne les séquences qui sont identiques à une région d'une séquence plus longue. 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

11. Toute séquence de nucléotides doit être représentée par un seul brin de codage, dans le sens 5'-3' et de gauche à droite, ou de gauche à droite de manière à reproduire le sens 5'-3'. Les désignations 5' et 3' ou toute autre désignation similaire ne doivent pas être incluses dans la séquence. Toute séquence de nucléotides représentée par deux brins de codage et divulguée par énumération des résidus des deux brins doit être représentée sous la forme :

a) 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

b) 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.

12. Aux fins de la présente norme, le premier nucléotide présenté dans la séquence correspond à la position de résidu numéro 1. Lorsque des séquences de nucléotides présentent une configuration circulaire, le déposant doit choisir le nucléotide à la position de résidu numéro 1. La numérotation est continue sur l'ensemble de la séquence dans le sens 5'-3', ou dans le sens qui reproduit le sens 5'-3'. Le dernier numéro de position de résidu doit correspondre au nombre de nucléotides de la séquence.

13. 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.

14. 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 dans le tableau de caractéristiques au sens du paragraphe 19.

15. Lorsqu'il convient d'employer un symbole ambigu (représentant deux nucléotides possibles ou plus), il faut choisir le symbole le plus restrictif indiqué à l'annexe I (voir section 1, tableau 1). Si par exemple un nucléotide dans une position quelconque pouvait être désigné 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 dans le tableau de caractéristiques. 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 dans le tableau de caractéristiques au sens des paragraphes 16, 17, 21 ou 93 à 96. On trouvera des détails sur la représentation des variantes de séquence, par exemple des alternatives, des suppressions, des adjonctions ou des remplacements, aux paragraphes 93 à 100.

16. Les nucléotides modifiés doivent être représentés dans la séquence comme les nucléotides non modifiés correspondants, 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), c'est-à-dire un nucléotide "other", par exemple un nucléotide n'existant pas à l'état naturel, doit être représenté par le symbole "n". Le symbole "n" est l'équivalent d'un seul résidu.

17. Tout nucléotide modifié doit être accompagné d'une description supplémentaire dans le tableau de caractéristiques (voir les paragraphes 60 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 du nucléotide modifié doit être indiqué dans un qualificateur de type "note". Pour un listage d'autres nucléotides modifiés, la valeur qualificative "OTHER" peut être utilisée conjointement avec un qualificateur de type "note" supplémentaire (voir les paragraphes 97 et 98). 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.

18. Une séquence de nucléotides comprenant une ou plusieurs régions de nucléotides modifiés consécutifs partageant la même fraction squelette (voir le paragraphe 3.g)i)2)) doit être accompagnée d'une description supplémentaire dans le tableau de caractéristiques comme indiqué au paragraphe 17. Les nucléotides modifiés de chacune de ces régions peuvent faire l'objet d'une description commune au moyen d'un seul élément INSDFeature comme indiqué au paragraphe 22. Le nom chimique complet non abrégé le plus restrictif qui englobe tous les nucléotides modifiés de la série ou une liste des noms chimiques de tous les nucléotides de la série doit être indiqué sous forme de valeur dans le qualificateur de type "note". Par exemple, la séquence d'un acide nucléique à glycol contenant des bases azotées "a", "c", "g", ou "t" peut comporter un qualificateur de type "note" dont la valeur est "2,3-dihydroxypropyl nuclosides". Cette même séquence peut également comporter un qualificateur de type "note" dont la valeur est "2,3-dihydroxypropyladenine, 2,3-dihydroxypropylthymine, 2,3-dihydroxypropylguanine, or 2,3-dihydroxypropylcytosine". Lorsqu'un nucléotide modifié de la région comporte une modification supplémentaire, celui-ci doit également être accompagné d'une description supplémentaire dans le tableau de caractéristiques comme indiqué au paragraphe 17.

19. 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".

20. Les exemples ci-après illustrent la manière dont des nucléotides modifiés doivent être représentés pour être conformes aux paragraphes 16 à 18 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>
```

Exemple 3 : Séquence de nucléotides composée de nucléotides modifiés visés par le paragraphe 3.g)i)2) avec deux nucléotides individuels comportant une modification supplémentaire

```
<INSDFeature>
 <INSDFeature_key>modified_base</INSDFeature_key>
 <INSDFeature_location>1..954</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>2,3-dihydroxypropyl  nucleosides</INSDQualifier_value>
     </INSDQualifier>
 </INSDFeature_quals>
</INSDFeature>
<INSDFeature>
 <INSDFeature_key>modified_base</INSDFeature_key>
 <INSDFeature_location>439</INSDFeature_location>
 <INSDFeature_quals>
     <INSDQualifier>
         <INSDQualifier_name>mod_base</INSDQualifier_name>
         <INSDQualifier_value>i</INSDQualifier_value>
     </INSDQualifier>
 </INSDFeature_quals>
</INSDFeature>
<INSDFeature>
 <INSDFeature_key>modified_base</INSDFeature_key>
 <INSDFeature_location>684</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>
```

21. 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.

22. 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 d'un seul élément INSDFeature avec la syntaxe "x..y" dans le descripteur d'emplacement de l'élément `INSDFeature_location` (voir les paragraphes 64 à 71). On trouvera des détails sur la représentation des variantes de séquence, par exemple des alternatives, des suppressions, des adjonctions ou des remplacements, aux paragraphes 93 à 100.

23. L'exemple ci-après illustre la représentation d'une région de nucléotides modifiés faisant l'objet de la même description au sens du paragraphe 22 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

24. Les acides aminés d'une séquence d'acides aminés doivent être représentés dans le sens amino-carboxy et de gauche à droite. Les groupes amino et carboxy ne doivent pas être représentés dans la séquence.

25. Aux fins de la présente norme, le premier acide aminé de la séquence est dans la position de résidu numéro 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. Lorsqu'une séquence d'acides aminés présente une configuration circulaire et que la séquence se compose essentiellement de résidus d'acides aminés reliés par des liaisons peptidiques, c'est-à-dire que la séquence ne comporte aucune terminaison amino ou carboxy, le déposant doit choisir l'acide aminé à la position de résidu numéro 1. La numérotation est continue sur l'ensemble de la séquence dans le sens amino-carboxy.

26. 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.

27. Lorsqu'il convient d'employer un symbole ambigu (représentant deux acides aminés possibles ou plus), il faut choisir le symbole le plus restrictif indiqué dans l'annexe I (voir section 3, tableau 3). 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 dans le tableau de caractéristiques. Le symbole "X" ne peut être employé que pour représenter un acide aminé. Il peut représenter un seul acide aminé modifié ou "inconnu" s'il est accompagné d'une description supplémentaire dans le tableau de caractéristiques, p. ex. au sens des paragraphes 29, 30, 32 ou 93 à 98. On trouvera des détails sur la représentation des variantes de séquence, par exemple des alternatives, des suppressions, des adjonctions ou des remplacements, aux paragraphes 93 à 100.

28. Les séquences d'acides aminés divulguées séparées par des symboles internes de fin (représentés par exemple par "Ter", l'astérisque "*", le point "." ou un espace blanc) doivent être ajoutées 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 7. Chacune de ces séquences distinctes doit disposer de son propre numéro d'identification de séquence. Les symboles de fin et les espaces blancs ne doivent pas être ajoutés dans les séquences figurant dans un listage (voir le paragraphe 57).

29. Les acides aminés modifiés, y compris les acides aminés D, 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), c'est-à-dire un acide aminé "autre", doit être représenté par le symbole "X". Ce symbole "X" n'est l'équivalent que d'un seul résidu.

30. Tout acide aminé modifié doit être accompagné d'une description supplémentaire dans le tableau de caractéristiques (voir les paragraphes 60 et suivants). Le cas échéant, les clés de caractérisation "CARBOHYD" ou "LIPID" devraient être utilisées avec le qualificateur "note". Il faudrait employer la clé de caractérisation "MOD_RES" et le qualificateur "note" pour les acides aminés autres modifiés après traduction; autrement, la clé de caractérisation "SITE" ainsi que le qualificateur "note" doivent être utilisés. 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.

31. Les exemples ci-après illustrent la manière dont des acides aminés modifiés doivent être représentés pour être conformes au paragraphe 30 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>3Hyp</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 : acide aminé D

```
<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>
```

32. Tout acide aminé "inconnu" doit être représenté par le symbole "X" dans la séquence. Tout acide aminé "inconnu" désigné par "X" doit être accompagné d'une description supplémentaire dans le tableau de caractéristiques (voir les paragraphes 60 et suivants) comportant la clé de caractérisation "UNSURE" et éventuellement le qualificateur "note". Le symbole "X" ne peut être l'équivalent que d'un seul résidu.

33. L'exemple ci-après illustre la manière dont un acide aminé "inconnu" doit être représenté pour être conforme au paragraphe 32 ci-dessus :

```
<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>
```

34. 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 64 à 70). On trouvera des détails sur la représentation des variantes de séquence, par exemple des alternatives, des suppressions, des adjonctions ou des remplacements, aux paragraphes 93 à 100.

35. 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 et doit disposer de son propre numéro d'identification de séquence.

36. Toute séquence qui contient des régions de résidus définis de manière spécifique et séparés par une ou plusieurs régions de résidus "n" ou "X" contigus (voir respectivement les paragraphes 15 et 27), et pour laquelle le nombre exact de résidus "n" ou "X" dans chaque région est divulgué, doit être ajoutée au listage des séquences comme une séquence et doit disposer de son propre numéro d'identification de séquence.

37. Toute séquence qui contient des régions de résidus définis de manière spécifique et séparés par une ou plusieurs brèches composées d'un nombre inconnu ou non divulgué de résidus ne doit pas être représentée dans le listage des séquences comme une séquence unique. Chaque région de résidus définis de manière spécifique visée par le paragraphe 7 doit être insérée dans le listage des séquences comme une séquence distincte et doit disposer de son propre numéro d'identification de séquence.

STRUCTURE DU LISTAGE DE SÉQUENCES EN XML

38. En application du paragraphe 6 ci-dessus, l'instance XML d'un fichier de listage des séquences conforme à la présente norme se compose des éléments suivants :

a) une partie consacrée aux informations générales, qui contient des informations sur la demande de brevet à laquelle se rapporte le listage des séquences; et

b) 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.

39. 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 (DTD) pour le listage des séquences".

a) La première ligne de l'instance XML doit contenir la déclaration du format XML :

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

b) La deuxième ligne de l'instance XML doit contenir une déclaration de type de document (DOCTYPE) :

```
<!DOCTYPE ST26SequenceListing PUBLIC "-//WIPO//DTD Sequence Listing 1.~~3~~//EN"
"ST26SequenceListing_V1_~~3~~.dtd">.
```

40. 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 :

a) les informations figurant dans les éléments `ApplicantName`, `InventorName` et `InventionTitle` dans la partie consacrée aux informations générales et l'élément `NonEnglishQualifier_value` dans la partie consacrée aux données sur les séquences peuvent comporter n'importe quel caractère Unicode valide indiqué dans la spécification XML 1.0 à l'exception des points de codes de contrôle Unicode 0000-001F et 007F-009F. Les caractères réservés ", &, ', <, et > (points de codes Unicode 0022, 0026, 0027, 003C et 003E respectivement) doivent être remplacés selon la méthode décrite au paragraphe 41; et

b) les informations figurant dans tous les autres éléments et attributs de la partie consacrée aux informations générales et dans tous les autres éléments et attributs 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 (c'est-à-dire qu'ils se limitent aux points de codes Unicode 0020 à 007E – voir l'annexe IV). Les caractères réservés ", &, ', <, et > (points de codes Unicode 0022, 0026, 0027, 003C et 003E respectivement) doivent être remplacés comme indiqué au paragraphe 41.

41. Dans l'instance XML d'un listage des séquences, les références sous forme de caractères numériques[1] ne doivent pas être utilisées et 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;

[1] Une référence sous forme de caractère numérique fait référence à un caractère selon son Jeu de caractères codés/point de code Unicode et utilise le format : "&#nnnn;" ou "&#xhhhh;", dans lequel "nnnn" est le point de code sous forme décimale et "hhhh" est le point de code sous forme hexadécimale.

‘	'

On trouvera un exemple au paragraphe 71. Les seules références d'entités de caractères autorisées sont les entités prédéfinies prévues dans ce paragraphe.

42. 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 (à l'exception de ce qui est prévu au paragraphe 97 concernant la représentation d'une suppression dans une séquence dans la valeur du qualificateur "replace").

Élément racine

43. 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_3".	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
`originalFreeTextLanguageCode`	Le code de langue (voir la référence au paragraphe 9 à la norme ISO 639-1:2002) pour la langue originale unique dans laquelle les qualificateurs de texte libre dépendant de la langue ont été établis.	Facultatif
`nonEnglishFreeTextLanguageCode`	Le code de langue (voir la référence au paragraphe 9 à la norme ISO 639-1 :2002) pour les éléments `NonEnglishQualifier_value`	Obligatoire lorsqu'un élément `NonEnglishQualifier_value` est présent dans le listage des séquences

44. 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_3" fileName="US11-405455-SEQL.xml"
softwareName="Wipo Sequence" softwareVersion="1.0" productionDate="2022-05-10"
originalFreeTextLanguageCode="de" nonEnglishFreeTextLanguageCode="fr">
  {...}*
</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

45. 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`	Numéro de la demande fourni 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 revendication de priorité la plus ancienne (contient également les éléments `IPOfficeCode`, `ApplicationNumberText` et `FilingDate`, voir `ApplicationIdentification` ci-dessus)	Obligatoire si une priorité est revendiquée.

Élément	Description	Obligatoire/ Facultatif
`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 devrait comporter 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 10).	Obligatoire

46. 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. 3//EN"
"ST26SequenceListing_V1_3.dtd">
<ST26SequenceListing dtdVersion="V1_3" fileName="Invention_SEQL.xml"
softwareName="WIPO Sequence" softwareVersion="1.0" productionDate="2022-05-10"
originalFreeTextLanguageCode="en" nonEnglishFreeTextLanguageCode="ja">
<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. 3//EN"
"ST26SequenceListing_V1_3.dtd">
<ST26SequenceListing dtdVersion="V1_3" fileName="Invention_SEQL.xml"
softwareName="WIPO Sequence" softwareVersion="1.0" productionDate="2022-05-10"
originalFreeTextLanguageCode="en" nonEnglishFreeTextLanguageCode="ja">
 <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.
```

47. 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 la référence à la norme ISO 639-1 :2002 au paragraphe 9) 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`.

48. 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 la référence à la norme ISO 639-1 :2002 au paragraphe 9) doit être indiqué dans l'attribut `languageCode` de l'élément.

49. 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 languageCodeja">特許 太郎</InventorName>
<InventorNameLatin>Taro Tokkyo</InventorNameLatin>
<InventionTitle  languageCode="ja">efg タンパク質をコードするマウス abcd-1 遺伝子
</InventionTitle>
<InventionTitle languageCode="en">Mus musculus abcd-1 gene for efg
protein</InventionTitle>
<InventionTitle languageCode="fr">Gène abcd-1 de Mus musculus pour protéine
efg</InventionTitle>
```

Partie consacrée aux données sur les séquences

50. 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.

51. Chaque élément `SequenceData` doit avoir un attribut obligatoire `sequenceIDNumber` contenant le numéro d'identification de la séquence (voir le paragraphe 10), par exemple :

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

52. 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"

53. 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>
```

54. L'élément `INSDSeq_moltype` doit divulguer le type de la molécule représentée. Pour les séquences de nucléotides, y compris les séquences d'analogues nucléotidiques, le type de molécule doit être ADN ou ARN. Pour les séquences d'acides aminés, le type de molécule doit être AA. (Cet élément est distinct du qualificateur "mol_type" mentionné aux paragraphes 55 et 84.) Par exemple :

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

55. Pour une séquence de nucléotides qui contient à la fois des segments d'ADN et d'ARN comprenant un ou plusieurs nucléotides, le type de molécule 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 segment 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 segment d'ADN ou d'ARN.

56. L'exemple ci-après illustre la description d'une séquence de nucléotides contenant à la fois des segments 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</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</INSDQualifier_value>
                    </INSDQualifier>
                </INSDFeature_quals>
            </INSDFeature>
        </INSDSeq_feature-table>
    <INSDSeq_sequence>cgacccacgcgtccgaggaaccaaccatcacgtttgaggacttcgtgaaggaattggataatacccgt
    ccctaccaaaatggcgagcgccgactcattgctcctcgtaccgtcgagcggc</INSDSeq_sequence>
    </INSDSeq>
```

57. L'élément `INSDSeq_sequence` doit divulguer la séquence. Seuls les symboles adéquats indiqués dans l'annexe I (voir section 1, tableau 1 et section 3, tableau 3) doivent figurer dans la séquence. La séquence ne doit pas contenir de chiffres, de signes de ponctuation ou d'espaces blancs.

58. Une séquence délibérément omise doit figurer dans le listage des séquences et être représentée de la manière suivante :

a) l'élément `SequenceData` et son attribut `sequenceIDNumber`, qui prend pour valeur le numéro d'identification de la séquence omise;

b) les éléments `INSDSeq_length`, `INSDSeq_moltype`, `INSDSeq_division`, qui sont présents mais ne contiennent aucune valeur;

c) l'élément `INSDSeq_feature-table` ne doit pas être indiqué; et

d) l'élément `INSDSeq_sequence`, qui prend la valeur "000".

59. L'exemple ci-après illustre la manière dont une séquence délibérément omise doit être représentée 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

60. 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`.

61. 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 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

62. 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

63. La clé de caractérisation "source" est obligatoire pour toutes les séquences de nucléotides et pour toutes les séquences d'acides aminés, sauf s'il s'agit d'une séquence délibérément omise. Chaque séquence doit comporter une clé "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.

Emplacement de la caractéristique

64. 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`. Les séquences d'acides aminés doivent contenir un seul descripteur d'emplacement dans l'élément `INSDFeature_location`. Les séquences de nucléotides peuvent comprendre plusieurs descripteurs dans l'élément `INSDFeature_location` lorsqu'un ou plusieurs descripteurs d'emplacement (voir les paragraphes 67 à 70) sont utilisé conjointement.

65. Le descripteur d'emplacement peut être un numéro de résidu unique, 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. 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`. Pour les séquences de nucléotides uniquement, un descripteur d'emplacement peut être un site entre deux numéros de résidus adjacents. 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 de nucléotides (voir les paragraphes 67 à 70).

66. 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 positifset inférieurs ou égaux à la longueur de la séquence dans l'élément `INSDSeq_sequence`, et où x est inférieur à y.

a) Descripteurs d'emplacement pour des séquences de nucléotides et d'acides aminés :

Type de descripteurs d'emplacement	Syntaxe	Description
Numéro de résidu unique	`x`	Désigne un résidu unique dans la séquence.
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`
	`<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.

b) Descripteurs d'emplacement pour les séquences de nucléotides uniquement :

Type de descripteurs d'emplacement	Syntaxe	Description
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.

c) Descripteurs d'emplacement pour les séquences d'acides aminés uniquement :

Type de descripteurs d'emplacement	Syntaxe	Description
Numéros de résidus reliés par une liaison intrachaîne	x..y	Désigne des acides aminés reliés par une liaison intrachaîne lorsqu'utilisés avec une caractéristique indiquant une liaison intrachaîne, telles que "CROSSLNK" ou "DISULFID".

67. L'élément `INSDFeature_location` des séquences de nucléotides peut contenir plusieurs opérateurs d'emplacement. 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. Les opérateurs d'emplacement peuvent être utilisés pour les nucléotides uniquement.

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.
`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', ou dans le sens qui reproduit le sens 5'-3'.

68. 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 doivent pas 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).

69. L'opérateur d'emplacement "complement" 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.

70. Les exemples ci-après illustrent des emplacements de caractéristiques au sens des paragraphes 64 à 69 ci-dessus :

a) emplacements pour les séquences de nucléotides et d'acides aminés :

Exemple d'emplacement	Description
467	Désigne le résidu 467 de la séquence.
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.

Exemple d'emplacement	Description
1..>888	Indique que la caractéristique commence au premier résidu de la séquence et continue au-delà du résidu 888.

b) emplacements pour les séquences de nucléotides uniquement :

Exemple d'emplacement	Description
123^124	Désigne un site entre les résidus 123 et 124
join(12..78,134..202)	Indique que les régions 12 à 78 et 134 à 202 devraient être jointes pour constituer une séquence contiguë
complement(34..126)	Commence au nucléotide complémentaire au nucléotide 126 et finit au nucléotide complémentaire au nucléotide 34 (la caractéristique est située sur le brin complémentaire au brin présenté).
complement(join(2691..4571, 4918..5163))	Joint les nucléotides 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).

c) emplacement pour les séquences d'acides aminés uniquement :

Exemple d'emplacement	Description
340…565	Indique que les acides aminés aux positions 340 et 565 sont reliés par une liaison intrachaîne lorsqu'utilisés avec une caractéristique qui indique une liaison intrachaîne, telle que "CROSSLNK" ou "DISULFID"

71. 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

72. 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 :

a) du texte libre (voir les paragraphes 85 à 87);

b) un vocabulaire contrôlé ou l'énumération de valeurs (p. ex. un nombre ou une date); et

c) des séquences.

73. 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 séquence de nucléotides, et la section 8 contient la liste complète des qualificateurs et la définition du format de leurs valeurs, le cas échéant, pour la clé de caractérisation de chaque séquence d'acides aminés.

74. Toute séquence prévue au paragraphe 7 qui est indiquée à titre de valeur d'un qualificateur doit figurer de manière distincte dans le listage des séquences et doit disposer de son propre numéro d'identification de séquence (voir le paragraphe 10).

Qualificateurs de caractéristiques obligatoires

75. Une clé de caractérisation obligatoire, en l'occurrence "source" pour les séquences de nucléotides et les séquences d'acides aminés, doit être accompagnée de deux qualificateurs obligatoires, "organism" et "mol_type". Certaines clés de caractérisation facultatives nécessitent aussi des qualificateurs obligatoires.

Éléments des qualificateurs

76. L'élément `INSDFeature_quals` se compose d'un ou plusieurs éléments `INSDQualifier`. Chaque élément `INSDQualifier` représente un seul qualificateur et se compose de trois éléments subordonnés et d'un attribut facultatif, de la manière suivante :

Élément/Attribut	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) et composé des caractères indiqués au paragraphe 40.b)	Obligatoire si indiqué (voir le paragraphe 87 et l'annexe I, sections 6 et 8)
NonEnglishQualifier_value	Valeur du qualificateur, le cas échéant, au format indiqué (voir l'annexe I, sections 6 et 8) et composé des caractères indiqués au paragraphe 40.a)	Obligatoire si indiqué (voir le paragraphe 87 et l'annexe I, sections 6 et 8)
id	Un qualificateur dont la valeur en texte libre dépend de la langue peut être identifié de manière unique en utilisant l'attribut XML facultatif 'id' dans l'élément INSDQualifier (voir le paragraphe 87.d)). La valeur de l'attribut 'id' doit commencer par la lettre 'q' et se poursuivre par tout nombre entier positif. La valeur d'un attribut 'id' doit être unique pour un élément INSDQualifier, c'est-à-dire que la valeur de l'attribut ne doit être utilisée qu'une seule fois dans un fichier de listage des séquences.	Facultatif

77. 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. Les indications d'organisme doivent être choisies parmi les éléments d'une taxonomie.

78. 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 les séquences d'acides aminés, mais ce nom ne doit pas être employé comme valeur du qualificateur d'organisme.

79. Les exemples suivants illustrent l'organisme source d'une séquence conformément aux paragraphes 77 et 78 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 d'acides aminés

```
<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>
```

80. 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>
```

81. Si la séquence existe à l'état naturel, 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". Toute information taxonomique connue doit être indiquée 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>
```

82. 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 (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>
```

83. 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>orgnism</INSDQualifier_name>
                <INSDQualifier_value>synthetic construct</INSDQualifier_value>
            </INSDQualifier>
            <INSDQualifier>
                <INSDQualifier_name>MOL_TYPEmol_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>
```

84. Le qualificateur "mol_type" pour les séquences de nucléotides (voir annexe I, section 6) et le qualificateur "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 :

a) 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";

b) pour des séquences d'acides aminés, la valeur du qualificateur "mol_type" est "protein".

Texte libre

85. Le texte libre, comme indiqué au paragraphe 3 (n), est un format de valeur autorisé pour certains qualificateurs. Il s'agit d'un texte descriptif qui se présente sous forme de segments de phrases ou dans tout autre format indiqué (comme indiqué à l'annexe I).

86. 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 autre que le qualificateur "translation", le texte libre ne peut compter plus de 1000 caractères.

87. Le texte libre dépendant de la langue, comme indiqué au paragraphe 3 (o), est la valeur de texte libre de certains qualificateurs qui est dépendante de la langue et peut nécessiter une traduction aux fins des procédures internationales, nationales ou régionales. Les qualificateurs pour les séquences de nucléotides avec un format de valeur de texte libre dépendant de la langue sont indiqués à l'annexe I, section 6 et tableau 5. Les qualificateurs pour les séquences d'acides aminés avec un format de valeur de texte libre dépendant de la langue sont indiqués à l'annexe I, section 8 et tableau 6.

a) Le texte libre dépendant de la langue doit être présenté dans l'élément `INSDQualifier_value` en anglais, ou dans l'élément `NonEnglishQualifier_value` dans une langue autre que l'anglais, ou dans les deux. Il convient de noter que si le nom d'un organisme est constitué par un nom de genre ou d'espèce en latin, aucune traduction n'est requise. Les termes techniques et les noms propres ayant pour origine des mots dans une langue autre que l'anglais et qui sont utilisés à l'échelle internationale sont considérés comme de l'anglais aux fins de la valeur de l'élément `INSDQualifier_value` (par exemple "in vitro" ou "in vivo").

b) Si un élément `NonEnglishQualifier_value` est présent dans un listage des séquences, le code de langue approprié (voir la référence au paragraphe 9 à la norme ISO 639-1:2002) doit être indiqué dans l'attribut `nonEnglishFreeTextLanguageCode` dans l'élément racine (voir le paragraphe 43). Tous les éléments `NonEnglishQualifier_value` dans un listage des séquences unique doivent avoir des valeurs dans la langue indiquée dans l'attribut `nonEnglishFreeTextLanguageCode`. L'élément `nonEnglishFreeTextLanguageCode` est autorisé uniquement pour les qualificateurs qui ont un format de valeur de texte libre dépendant de la langue.

c) Lorsque `NonEnglishQualifier_value` et `INSDQualifier_value` sont tous deux présents pour un seul qualificateur, les informations contenues dans les deux éléments doivent être équivalentes. Une des condictions ci-après doit être satisfaite : `NonEnglishQualifier_value` contient une traduction de la valeur `INSDQualifier_value`; ou `INSDQualifier_value` contient une traduction de la valeur `NonEnglishQualifier_value`; ou les deux éléments contiennent une traduction de la valeur du qualificateur à partir de la langue indiquée dans l'attribut `originalFreeTextLanguageCode` (voir le paragraphe 43).

d) Pour les qualificateurs avec une valeur de texte libre dépendant de la langue, l'élément `INSDQualifier_value` peut comporter un attribut facultatif `id`. La valeur de cet attribut doit être dans le format "q" suivi d'un nombre entier positif, par exemple "q23", et doit être unique pour un élément `INSDQualifier_value`, c'est-à-dire que la valeur de l'attribut ne doit être utilisée qu'une seule fois dans un fichier de listage des séquences.

88. Les exemples ci-après illustrent la présentation de texte libre dépendant de la langue indiquée au paragraphe 87.

Exemple 1 : texte libre dépendant de la langue dans un élément `INSDQualifier_value` :

```
<INSDFeature>
      <INSDFeature_key>regulatory</INSDFeature_key>
      <INSDFeature_location>1..60</INSDFeature_location>
      <INSDFeature_quals>
        <INSDQualifier id="q1">
               <INSDQualifier_name>function</INSDQualifier_name>
               <INSDQualifier_value>binds to regulatory protein Est3</INSDQualifier_value>
        </INSDQualifier>
      </INSDFeature_quals>
</INSDFeature>
```

Exemple 2 : texte libre dépendant de la langue dans un élément `INSDQualifier_value` et dans un élément `NonEnglishQualifier_value` :

```
<INSDFeature>
        <INSDFeature_key>ACT_SITE</INSDFeature_key>
        <INSDFeature_location>51..64</INSDFeature_location>
        <INSDFeature_quals>
            <INSDQualifier id="q45">
                <INSDQualifier_name>note</INSDQualifier_name>
                <INSDQualifier_value>cleaves carbohydrate chain</INSDQualifier_value>
                <NonEnglishQualifier_value>clive la chaîne glucidique
                </NonEnglishQualifier_value>
            </INSDQualifier>
        </INSDFeature_quals>
</INSDFeature>
```

Exemple 3 : texte libre dépendant de la langue dans un élément `NonEnglishQualifier_value` :

```
<INSDFeature>
        <INSDFeature_key>ACT_SITE</INSDFeature_key>
        <INSDFeature_location>51..64</INSDFeature_location>
        <INSDFeature_quals>
            <INSDQualifier id="q45">
                <INSDQualifier_name>note</INSDQualifier_name>
                <INSDQualifier_value>cleaves carbohydrate chain</INSDQualifier_value>
                <NonEnglishQualifier_value>clive la chaîne glucidique
                </NonEnglishQualifier_value>
            </INSDQualifier>
        </INSDFeature_quals>
</INSDFeature>
```

Séquences de codage

89. 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'emplacement de la caractéristique "CDS" dans l'élément obligatoire `INSDFeature_location` doit indiquer le codon d'arrêt.

90. 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 7).

91. Le qualificateur "transl_except" doit être employé en association avec la clé de caractérisation "CDS" et le qualificateur "translation" doit être employé pour désigner un codon codant pour la pyrrolysine ou la sélénocystéine.

92. Toute séquence d'acides aminés codée selon la séquence de codage et divulguée dans un qualificateur de type "translation" visé par le paragraphe 7 doit figurer dans le listage des séquences et doit disposer de son propre numéro d'identification de séquence. Le numéro d'identification de séquence attribué à la séquence d'acides aminés 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 d'acides aminés doit être identique à celui de sa séquence de codage, par exemple :

```
          <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>
     MLVHLERTTIMFDFSSLINLPLIWGLLIAIAVLLYILMDGFDLGIGILLPFAPSDKCRDHMISSIAPFWDGNETWLVLGGGGLFAA
     FPLAYSILMPAFYIPIIIMLLGLIVRGVSFEFRFKAEGKYRRLWDYAFHFGSLGAAFCQGMILGAFIHGVEVNGRNFSGGQLM
                  </INSDQualifier_value>
              </INSDQualifier>
              <INSDQualifier>
                      <INSDQualifier_name>protein_id</INSDQualifier_name>
                  <INSDQualifier_value>89</INSDQualifier_value>
              </INSDQualifier>
          </INSDFeature_quals>
      </INSDFeature>
```

Variantes

93. Toute séquence primaire et toute variante de cette séquence, chacune d'elles étant divulguée par énumération de ses résidus et visée par le paragraphe 7, doit figurer dans le listage des séquences et doit disposer de son propre numéro d'identification de séquence.

94. Toute séquence variante, divulguée comme une séquence unique avec des résidus alternatifs énumérés à une ou plusieurs positions, doit figurer dans le listage des séquences et devrait être représentée par une séquence unique, les résidus alternatifs énumérés étant représentés par le symbole le plus restrictif (voir les paragraphes 15 et 27).

95. Toute séquence variante, divulguée uniquement par référence à un ou plusieurs suppressions, adjonctions ou remplacements effectués dans une séquence primaire figurant dans le listage des séquences, doit figurer dans le listage des séquences. Si tel est le cas, cette séquence variante :

a) peut être représentée 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;

b) devrait être représentée en tant que séquence distincte et devrait disposer de 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

c) doit être représentée en tant que séquence distincte et doit disposer de 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 86).

96. Le tableau ci-dessous indique le bon usage des clés de caractérisation et des qualificateurs pour des variantes de séquences d'acides nucléiques et d'acides aminés :

Type de séquence	Clé de caractérisation	Qualificateur	Usage
Acide nucléique	variation	replace or note	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 or note	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

97. L'annotation d'une séquence 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. La valeur du qualificateur "replace" doit correspondre uniquement à un nucléotide alternatif unique ou à une séquence de nucléotides alternatifs uniques représenté à l'aide des symboles indiqués dans la section 1 du tableau 1, ou vide. Un listage des résidus alternatifs peut être indiqué dans le qualificateur "note". Il convient en particulier d'indiquer un listage d'acides aminés alternatifs dans le qualificateur "note" si "X" est employé dans une séquence et représente une valeur autre que "l'équivalent de 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'" (voir le paragraphe 27). Toute suppression doit être représentée par une valeur de qualificateur vide pour le qualificateur "replace" ou par une indication dans le qualificateur "note" selon laquelle le résidu peut être supprimé. Tout résidu ajouté ou remplacé doit être indiqué dans le qualificateur "replace" ou "note". La valeur du qualificateur "replace" et "note" est un texte libre qui ne doit pas dépasser 1000 caractères, conformément au paragraphe 86. Pour les séquences visées par le paragraphe 7 qui sont présentées à titre d'adjonction ou de remplacement de la valeur d'un qualificateur, se reporter au paragraphe 100.

98. 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. Pour le qualificateur "note", 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. Les résidus modifiés doivent être accompagnés d'une description supplémentaire dans le tableau de caractéristiques comme prévu aux paragraphes 17 ou 30.

99. Les exemples ci-après illustrent la manière de représenter des variantes pour qu'elles soient conformes aux paragraphes 95 à 98 ci-dessus :

Exemple 1 : Clé de caractérisation "misc_difference" pour des nucléotides alternatifs énumérés. Le "n" à la position 53 de la séquence peut être un nucléotide alternatif parmi cinq nucléotides alternatifs.

```
<INSDFeature>
 <INSDFeature_key>misc_difference</INSDFeature_key>
 <INSDFeature_location>53</INSDFeature_location>
 <INSDFeature_quals>
     <INSDQualifier>
         <INSDQualifier_name>note</INSDQualifier_name>
         <INSDQualifier_value>w, cmnm5s2u, mam5u, mcm5s2u, or p
</INSDQualifier_value>
     </INSDQualifier>
 </INSDFeature_quals>
</INSDFeature>
<INSDFeature>
 <INSDFeature_key>modified_base</INSDFeature_key>
 <INSDFeature_location>53</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>cmnm5s2u, mam5u, mcm5s2u, or p</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 “variation” pour un remplacement dans une séquence nucléotidique.

Une cytosine remplace le nucléotide indiqué 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 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 I, A, F, Y, alle, Melle, 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>
<INSDFeature>
 <INSDFeature_key>MOD_RES</INSDFeature_key>
 <INSDFeature_location>100</INSDFeature_location>
 <INSDFeature_quals>
     <INSDQualifier>
         <INSDQualifier_name>note</INSDQualifier_name>
         <INSDQualifier_value>aIle, MeIle, or Nle</INSDQualifier_value>
     </INSDQualifier>
 </INSDFeature_quals>
</INSDFeature>
```

Exemple 6 : 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>
```

100. Toute séquence visée par le paragraphe 7 qui est présentée à titre d'adjonction ou de remplacement dans la valeur d'un qualificateur pour une annotation de séquence primaire doit aussi figurer dans le listage des séquences et disposer de son propre numéro d'identification de séquence.

[L'annexe I suit]

ANNEXE I

VOCABULAIRE CONTROLE

Version 1.7

Révision approuvée pour approbation au Comité des normes de l'OMPI (CWS) à sa onzième session le 8 décembre 2023

TABLE DES MATIERES

SECTION 1 : LISTE DES NUCLÉOTIDES

Les symboles des bases nucléotidiques à utiliser dans les listages des séquences sont présentés dans le tableau 1. Lorsqu'il n'est pas accompagné d'une description supplémentaire, le symbole "t" désigne la thymine dans de l'ADN et l'uracile dans de l'ARN. Lorsqu'il convient d'employer un symbole ambigu (représentant deux bases nucléotidiques possibles ou plus), il faut choisir le symbole le plus restrictif. Si, par exemple, une base nucléotidique 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 "a ou c ou g ou t/u" s'il n'est pas accompagné d'une description supplémentaire.

Tableau 1 : Liste des symboles des nucléotides

Symbole	Définition
a	Adénine
c	Cytosine
g	Guanine
t	Thymine dans de l'ADN/uracile dans de l'ARN (t/u)
m	a ou c
r	a ou g
w	a ou t/u
s	c ou g
y	c ou t/u
k	g ou t/u
v	a ou c ou g; et non t/u
h	a ou c ou t/u; et non g
d	a ou g ou t/u; et non c
b	c ou g ou t/u; et non a
n	a ou c ou g ou t/u; "unknown" ou "other"

SECTION 2 : LISTE DES NUCLÉOTIDES MODIFIÉS

Les abréviations indiquées dans le tableau 2 sont les seules valeurs autorisées pour le qualificateur mod_base. Lorsqu'un nucléotide modifié particulier ne figure pas dans le tableau ci-après, il doit prendre pour valeur l'abréviation "OTHER". Si l'abréviation est "OTHER", le nom complet non abrégé de la base modifiée doit être indiqué dans un qualificateur du type "note". Les abréviations indiquées dans le tableau 2 ne doivent pas être employées dans la séquence elle-même.

Tableau 2 : Liste des nucléotides modifiés

Abréviation	Définition
ac4c	4-acetylcytidine
chm5u	5-(carboxyhydroxylmethyl)uridine
cm	2'-O-methylcytidine
cmnm5s2u	5-carboxymethylaminomethyl-2-thiouridine
cmnm5u	5-carboxymethylaminomethyluridine
dhu	Dihydrouridine
fm	2'-O-methylpseudouridine
gal q	beta-D-galactosylqueuosine
gm	2'-O-methylguanosine
i	Inosine
i6a	N6-isopentenyladenosine
m1a	1-methyladenosine
m1f	1-methylpseudouridine
m1g	1-methylguanosine
m1i	1-methylinosine
m22g	2,2-dimethylguanosine
m2a	2-methyladenosine
m2g	2-methylguanosine
m3c	3-methylcytidine

m4c	N4-methylcytosine
m5c	5-methylcytidine
m6a	N6-methyladenosine
m7g	7-methylguanosine
mam5u	5-methylaminomethyluridine
mam5s2u	5-methylaminomethyl-2-thiouridine
man q	beta-D-mannosylqueuosine
mcm5s2u	5-methoxycarbonylmethyl-2-thiouridine
mcm5u	5-methoxycarbonylmethyluridine
mo5u	5-methoxyuridine
ms2i6a	2-methylthio-N6-isopentenyladenosine
ms2t6a	N-((9-beta-D-ribofuranosyl-2-methylthiopurine-6-yl)carbamoyl)threonine
mt6a	N-((9-beta-D-ribofuranosylpurine-6-yl)N-methyl-carbamoyl)threonine
mv	uridine-5-oxoacetic acid-methylester
o5u	uridine-5-oxyacetic acid (v)
osyw	Wybutoxosine
p	Pseudouridine
q	Queuosine
s2c	2-thiocytidine
s2t	5-methyl-2-thiouridine
s2u	2-thiouridine
s4u	4-thiouridine
m5u	5-methyluridine
t6a	N-((9-beta-D-ribofuranosylpurine-6-yl)carbamoyl)threonine
tm	2'-O-methyl-5-methyluridine
um	2'-O-methyluridine
yw	Wybutosine
x	3-(3-amino-3-carboxypropyl)uridine, (acp3)u
OTHER	(nécessite un qualificateur “note”)

SECTION 3 : LISTE DES ACIDES AMINÉS

Les symboles des acides aminés à employer dans la séquence sont présentés dans le tableau 3. 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 à une position quelconque pouvait être un acide aspartique ou une asparagine, il faut employer le symbole "B" au lieu de "X". Le symbole "X" sera considéré comme l'équivalent de 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" s'il n'est pas accompagné d'une description supplémentaire.

Tableau 3 : Liste des symboles des acides aminés

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

SECTION 4 : LISTE DES ACIDES AMINÉS MODIFIÉS

Le tableau 4 indique les seules abréviations autorisées pour un acide aminé modifié dans le qualificateur obligatoire du type "note" pour les clés de caractérisation "MOD_RES" ou "SITE". La valeur du qualificateur du type "note" doit être soit une abréviation indiquée dans ce tableau, s'il y a lieu, soit le nom complet non abrégé de l'acide aminé modifié. Les abréviations (ou les noms complets) indiquées dans ce tableau ne doivent pas être employées dans la séquence elle-même.

Tableau 4 : Liste des acides aminés modifiés

Abréviation	Acide aminé modifié
Aad	2-Aminoadipic acid
bAad	3-Aminoadipic acid
bAla	beta-Alanine, beta-Aminoproprionic acid
Abu	2-Aminobutyric acid
4Abu	4-Aminobutyric acid, piperidinic acid
Acp	6-Aminocaproic acid
Ahe	2-Aminoheptanoic acid
Aib	2-Aminoisobutyric acid
bAib	3-Aminoisobutyric acid
Apm	2-Aminopimelic acid
Dbu	2,4-Diaminobutyric acid
Des	Desmosine
Dpm	2,2'-Diaminopimelic acid
Dpr	2,3-Diaminoproprionic acid
EtGly	N-Ethylglycine
EtAsn	N-Ethylasparagine
Hyl	Hydroxylysine
aHyl	allo-Hydroxylysine
3Hyp	3-Hydroxyproline
4Hyp	4-Hydroxyproline
Ide	Isodesmosine
aIle	allo-Isoleucine
MeGly	N-Methylglycine, sarcosine
MeIle	N-Methylisoleucine
MeLys	6-N-Methyllysine
MeVal	N-Methylvaline
Nva	Norvaline
Nle	Norleucine
Orn	Ornithine

SECTION 5 : CLÉS DE CARACTÉRISATION POUR LES SÉQUENCES DE NUCLÉOTIDES

La présente section donne la liste des clés de caractérisation qui peuvent être employées pour les séquences de nucléotides, ainsi qu'une liste des qualificateurs obligatoires et facultatifs. Les clés de caractérisation sont présentées dans l'ordre alphabétique. Sauf indication contraire, elles peuvent être employées soit pour l'ADN, soit pour l'ARN sous "Molecule scope". Certaines Feature Keys peuvent être utilisées avec des séquences artificielles pour compléter le "organism scope" indiqué.

Les noms des clés de caractérisation doivent être employés dans l'instance XML du listage des séquences exactement comme ils apparaissent à la suite de "Feature key" dans les descriptions ci-après, à l'exception des clés de caractérisation 3'UTR et 5'UTR. On se reportera dans la description à "Comment" correspondant aux clés de caractérisation 3'UTR et 5'UTR.

5.1.	Feature Key	C_region
	Definition	constant region of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; includes one or more exons depending on the particular chain
	Optional qualifiers	allele gene gene_synonym map note product pseudo pseudogene standard_name
	Organism scope	eukaryotes

5.2.	Feature Key	CDS
	Definition	coding sequence; sequence of nucleotides that corresponds with the sequence of amino acids in a protein (location includes stop codon); feature may include amino acid conceptual translation
	Optional qualifiers	allele circular_RNA codon_start EC_number exception function gene gene_synonym map note number operon product protein_id pseudo pseudogene ribosomal_slippage standard_name translation transl_except transl_table trans_splicing
	Comment	codon_start qualifier has valid value of 1 or 2 or 3, indicating the offset at

		which the first complete codon of a coding feature can be found, relative to the first base of that feature; transl_table defines the genetic code table used if other than the Standard or universal genetic code table; genetic code exceptions outside the range of the specified tables are reported in transl_except qualifier; only one of the qualifiers translation, pseugogene or pseudo are permitted with a CDS feature key; when the translation qualifier is used, the protein_id qualifier is mandatory if the translation product contains four or more specifically defined amino acids
5.3.	Feature Key	centromere
	Definition	region of biological interest identified as a centromere and which has been experimentally characterized
	Optional qualifiers	note standard_name
	Comment	the centromere feature describes the interval of DNA that corresponds to a region where chromatids are held and a kinetochore is formed
5.4.	Feature Key	D-loop
	Definition	displacement loop; a region within mitochondrial DNA in which a short stretch of RNA is paired with one strand of DNA, displacing the original partner DNA strand in this region; also used to describe the displacement of a region of one strand of duplex DNA by a single stranded invader in the reaction catalyzed by RecA protein
	Optional qualifiers	allele gene gene_synonym map note
	Molecule scope	DNA
5.5.	Feature Key	D_segment
	Definition	Diversity segment of immunoglobulin heavy chain, and T-cell receptor beta chain
	Optional qualifiers	allele gene gene_synonym map note product pseudo pseudogene standard_name
	Organism scope	eukaryotes
5.6.	Feature Key	exon

Definition	region of genome that codes for portion of spliced mRNA,rRNA and tRNA; may contain 5'UTR, all CDSs and 3' UTR
Optional qualifiers	allele EC_number function gene gene_synonym map note number product pseudo pseudogene standard_name

5.7. Feature Key	gene
Definition	region of biological interest identified as a gene and for which a name has been assigned
Optional qualifiers	allele function gene gene_synonym map note operon product pseudo pseudogene phenotype standard_name trans_splicing
Comment	the gene feature describes the interval of DNA that corresponds to a genetic trait or phenotype; the feature is, by definition, not strictly bound to its positions at the ends; it is meant to represent a region where the gene is located.

5.8. Feature Key	iDNA
Definition	intervening DNA; DNA which is eliminated through any of several kinds of recombination
Optional qualifiers	allele function gene gene_synonym map note number standard_name
Molecule scope	DNA
Comment	e.g., in the somatic processing of immunoglobulin genes.

5.9.	Feature Key	intron
	Definition	a segment of DNA that is transcribed, but removed from within the transcript by splicing together the sequences (exons) on either side of it
	Optional qualifiers	allele function gene gene_synonym map note number pseudo pseudogene standard_name trans_splicing
5.10.	Feature Key	J_segment
	Definition	joining segment of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains
	Optional qualifiers	allele gene gene_synonym map note product pseudo pseudogene standard_name
	Organism scope	eukaryotes
5.11.	Feature Key	mat_peptide
	Definition	mature peptide or protein coding sequence; coding sequence for the mature or final peptide or protein product following post-translational modification; the location does not include the stop codon (unlike the corresponding CDS)
	Optional qualifiers	allele EC_number function gene gene_synonym map note product pseudo pseudogene standard_name

5.12.	Feature Key	misc_binding
	Definition	site in nucleic acid which covalently or non-covalently binds another moiety that cannot be described by any other binding key (primer_bind or protein_bind)
	Mandatory qualifiers	bound_moiety
	Optional qualifiers	allele function gene gene_synonym map note
	Comment	note that the regulatory feature key and regulatory_class qualifier with the value"ribosome_binding_site" must be used for describing ribosome binding sites

5.13.	Feature Key	misc_difference
	Definition	featured sequence differs from the presented sequence at this location and cannot be described by any other Difference key (variation, or modified_base)
	Optional qualifiers	allele clone compare gene gene_synonym map note phenotype replace standard_name
	Comment	the misc_difference feature key must be used to describe variability introduced artificially, e.g., by genetic manipulation or by chemical synthesis; use the replace qualifier to annotate a deletion, insertion, or substitution. The variation feature key must be used to describe naturally occurring genetic variability.

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

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

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

5.17.	Feature Key	misc_structure
	Definition	any secondary or tertiary nucleotide structure or conformation that cannot be described by other Structure keys (stem_loop and D-loop)
	Optional qualifiers	allele function gene gene_synonym map note standard_name

5.18.	Feature Key	mobile_element
	Definition	region of genome containing mobile elements
	Mandatory qualifiers	mobile_element_type
	Optional qualifiers	allele function gene gene_synonym map note rpt_family rpt_type standard_name

5.19.	Feature Key	modified_base
	Definition	the indicated nucleotide is a modified nucleotide and should be substituted for by the indicated molecule (given in the mod_base qualifier value)
	Mandatory qualifiers	mod_base
	Optional qualifiers	allele frequency gene gene_synonym map note
	Comment	value for the mandatory mod_base qualifier is limited to the restricted vocabulary for modified base abbreviations in Section 2 of this Annex.

5.20.	Feature Key	mRNA
	Definition	messenger RNA; includes 5' untranslated region (5'UTR), coding sequences (CDS, exon) and 3' untranslated region (3'UTR)
	Optional qualifiers	allele circular_RNAfunction gene gene_synonym map note operon product pseudo pseudogene standard_name trans_splicing

5.21.	.Feature Key	ncRNA
	Definition	a non-protein-coding gene, other than ribosomal RNA and transfer RNA, the functional molecule of which is the RNA transcript
	Mandatory qualifiers	ncRNA_class
	Optional qualifiers	allele function gene gene_synonym map note operon product pseudo pseudogene standard_name trans_splicing
	Comment	the ncRNA feature must not be used for ribosomal and transfer RNA annotation, for which the rRNA and tRNA feature keys must be used, respectively

5.22.	Feature Key	N_region
	Definition	extra nucleotides inserted between rearranged immunoglobulin segments
	Optional qualifiers	allele gene gene_synonym map note product pseudo pseudogene standard_name
	Organism scope	eukaryotes

5.23.	Feature Key	operon
	Definition	region containing polycistronic transcript including a cluster of genes that are under the control of the same regulatory sequences/promoter and in the same biological pathway
	Mandatory qualifiers	operon
	Optional qualifiers	allele function map note phenotype pseudo pseudogene standard_name

5.24.	Feature Key	oriT
	Definition	origin of transfer; region of a DNA molecule where transfer is initiated during the process of conjugation or mobilization
	Optional qualifiers	allele bound_moiety direction gene gene_synonym map note rpt_family rpt_type rpt_unit_range rpt_unit_seq standard_name
	Molecule Scope	DNA
	Comment	rep_origin must be used to describe origins of replication; direction qualifier has permitted values left, right, and both, however only left and right are valid when used in conjunction with the oriT feature; origins of transfer can be present in the chromosome; plasmids can contain multiple origins of transfer

5.25.	Feature Key	polyA_site
	Definition	site on an RNA transcript to which will be added adenine residues by post-transcriptional polyadenylation
	Optional qualifiers	allele gene gene_synonym map note
	Organism scope	eukaryotes and eukaryotic viruses

5.26.	Feature Key	precursor_RNA
	Definition	any RNA species that is not yet the mature RNA product; may include ncRNA, rRNA, tRNA, 5' untranslated region (5'UTR), coding sequences (CDS, exon), intervening sequences (intron) and 3' untranslated region (3'UTR)
	Optional qualifiers	allele function gene gene_synonym map note operon product standard_name trans_splicing
	Comment	used for RNA which may be the result of post-transcriptional processing; if the RNA in question is known not to have been processed, use the prim_transcript key

5.27.	Feature Key	prim_transcript
	Definition	primary (initial, unprocessed) transcript; may include ncRNA, rRNA, tRNA, 5' untranslated region (5'UTR), coding sequences (CDS, exon), intervening sequences (intron) and 3' untranslated region (3'UTR)
	Optional qualifiers	allele function gene gene_synonym map note operon standard_name

5.28.	Feature Key	primer_bind
	Definition	non-covalent primer binding site for initiation of replication, transcription, or reverse transcription; includes site(s) for synthetic e.g., PCR primer elements
	Optional qualifiers	allele gene gene_synonym map note standard_name
	Comment	used to annotate the site on a given sequence to which a primer molecule binds – not intended to represent the sequence of the primer molecule itself; since PCR reactions most often involve pairs of primers, a single primer_bind key may use the order(location,location) operator with two locations, or a pair of primer_bind keys may be used

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

	map note rpt_family rpt_type rpt_unit_range rpt_unit_seq satellite standard_name

5.33. Feature Key	rep_origin
Definition	origin of replication; starting site for duplication of nucleic acid to give two identical copies
Optional Qualifiers	allele direction function gene gene_synonym map note standard_name
Comment	direction qualifier has valid values: left, right, or both

5.34. Feature Key	rRNA
Definition	mature ribosomal RNA; RNA component of the ribonucleoprotein particle (ribosome) which assembles amino acids into proteins
Optional qualifiers	allele function gene gene_synonym map note operon product pseudo pseudogene standard_name
Comment	rRNA sizes should be annotated with the product qualifier

5.35. Feature Key	S_region
Definition	switch region of immunoglobulin heavy chains; involved in the rearrangement of heavy chain DNA leading to the expression of a different immunoglobulin class from the same B-cell
Optional qualifiers	allele gene gene_synonym map note product pseudo pseudogene standard_name
Organism scope	eukaryotes

5.36.	Feature Key	sig_peptide
	Definition	signal peptide coding sequence; coding sequence for an N-terminal domain of a secreted protein; this domain is involved in attaching nascent polypeptide to the membrane leader sequence
	Optional qualifiers	allele function gene gene_synonym map note product pseudo pseudogene standard_name

5.37.	Feature Key	source
	Definition	identifies the source of the sequence; this key is mandatory; every sequence will have a single source key spanning the entire sequence
	Mandatory qualifiers	organism mol_type
	Optional qualifiers	cell_line cell_type chromosome clone clone_lib collected_by collection_date cultivar dev_stage ecotype environmental_sample germline haplogroup haplotype host identified_by isolate isolation_source lab_host lat_lon macronuclear map mating_type note organelle PCR_primers plasmid pop_variant proviral rearranged segment serotype serovar sex strain sub_clone sub_species sub_strain tissue_lib tissue_type variety

	Molecule scope	any
5.38.	Feature Key	stem_loop
	Definition	hairpin; a double-helical region formed by base-pairing between adjacent (inverted) complementary sequences in a single strand of RNA or DNA
	Optional qualifiers	allele function gene gene_synonym map note operon standard_name
5.39.	Feature Key	STS
	Definition	sequence tagged site; short, single-copy DNA sequence that characterizes a mapping landmark on the genome and can be detected by PCR; a region of the genome can be mapped by determining the order of a series of STSs
	Optional qualifiers	allele gene gene_synonym map note standard_name
	Molecule scope	DNA
	Comment	STS location to include primer(s) in primer_bind key or primers
5.40.	Feature Key	telomere
	Definition	region of biological interest identified as a telomere and which has been experimentally characterized
	Optional qualifiers	note rpt_type rpt_unit_range rpt_unit_seq standard_name
	Comment	the telomere feature describes the interval of DNA that corresponds to a specific structure at the end of the linear eukaryotic chromosome which is required for the integrity and maintenance of the end; this region is unique compared to the rest of the chromosome and represents the physical end of the chromosome
5.41.	Feature Key	tmRNA
	Definition	transfer messenger RNA; tmRNA acts as a tRNA first, and then as an mRNA that encodes a peptide tag; the ribosome translates this mRNA region of tmRNA and attaches the encoded peptide tag to the C-terminus of the unfinished protein; this attached tag targets the protein for destruction or proteolysis
	Optional qualifiers	allele function gene gene_synonym map note product pseudo

		pseudogene standard_name tag_peptide
5.42.	Feature Key	transit_peptide
	Definition	transit peptide coding sequence; coding sequence for an N-terminal domain of a nuclear-encoded organellar protein; this domain is involved in post-translational import of the protein into the organelle
	Optional qualifiers	allele function gene gene_synonym map note product pseudo pseudogene standard_name
5.43.	Feature Key	tRNA
	Definition	mature transfer RNA, a small RNA molecule (75-85 bases long) that mediates the translation of a nucleic acid sequence into an amino acid sequence
	Optional qualifiers	allele circular_RNA anticodon function gene gene_synonym map note operon product pseudo pseudogene standard_name trans_splicing
5.44.	Feature Key	unsure
	Definition	a small region of sequenced bases, generally 10 or fewer in its length, which could not be confidently identified. Such a region might contain called bases (a, t, g, or c), or a mixture of called-bases and uncalled-bases ('n').
	Optional qualifiers	allele compare gene gene_synonym map note replace
	Comment	use the replace qualifier to annotate a deletion, insertion, or substitution.
5.45.	Feature Key	V_region
	Definition	variable region of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; codes for the variable amino terminal portion; can be composed of V_segments, D_segments, N_regions, and J_segments
	Optional qualifiers	allele gene gene_synonym

		map note product pseudo pseudogene standard_name
	Organism scope	eukaryotes

5.46.	Feature Key	V_segment
	Definition	variable segment of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; codes for most of the variable region (V_region) and the last few amino acids of the leader peptide
	Optional qualifiers	allele gene gene_synonym map note product pseudo pseudogene standard_name
	Organism scope	eukaryotes

5.47.	Feature Key	variation
	Definition	a related strain contains stable mutations from the same gene (e.g., RFLPs, polymorphisms, etc.) which differ from the presented sequence at this location (and possibly others)
	Optional qualifiers	allele compare frequency gene gene_synonym map note phenotype product replace standard_name
	Comment	used to describe alleles, RFLP's, and other naturally occurring mutations and polymorphisms; use the replace qualifier to annotate a deletion, insertion, or substitution; variability arising as a result of genetic manipulation (e.g., site directed mutagenesis) must be described with the misc_difference feature

5.48.	Feature Key	3'UTR
	Definition	1) region at the 3' end of a mature transcript (following the stop codon) that is not translated into a protein; 2) region at the 3' end of an RNA virus (following the last stop codon) that is not translated into a protein;
	Optional qualifiers	allele function gene gene_synonym map note standard_name

		trans_splicing
	Comment	The apostrophe character has special meaning in XML, and must be substituted with "'" in the value of an element. Thus "3'UTR" must be represented as "3'UTR" in the XML file, i.e., <INSDFeature_key>3'UTR</INSDFeature_key>.
5.49.	Feature Key	5'UTR
	Definition	1) region at the 5' end of a mature transcript (preceding the initiation codon) that is not translated into a protein; 2) region at the 5' end of an RNA virus (preceding the first initiation codon) that is not translated into a protein;
	Optional qualifiers	allele function gene gene_synonym map note standard_name trans_splicing
	Comment	The apostrophe character has special meaning in XML, and must be substituted with "'" in the value of an element. Thus "5'UTR" must be represented as "5'UTR" in the XML file, i.e., <INSDFeature_key>5'UTR</INSDFeature_key>.

SECTION 6 : QUALIFICATEURS POUR LES SÉQUENCES DE NUCLÉOTIDES

La présente section donne la liste des qualificateurs à employer pour les caractéristiques dans les séquences de nucléotides. Les qualificateurs sont présentés dans l'ordre alphabétique.

Lorsque le format de valeur est "none", il ne faut utiliser ni l'élément `INSDQualifier_value` ni l'élément `NonEnglishQualifier_value`.

Lorsque le format de valeur est du texte libre indiqué comme dépendant de la langue, un des éléments ci-après doit être utilisé :

1) l'élément `INSDQualifier_value`; ou

2) l'élément `NonEnglishQualifier_value`; ou

3) à la fois l'élément `INSDQualifier_value` et l'élément `NonEnglishQualifier_value`.

Lorsque le format de valeur est autre que "none" mais ne constitue pas du texte libre dépendant de la langue, il faut utiliser l'élément `INSDQualifier_value` et il ne faut pas utiliser l'élément `NonEnglishQualifier_value`.

N.B. : Toute valeur de qualificateur indiquée pour un qualificateur avec un format de valeur free text dépendant de la langue peut devoir être traduite aux fins des procédures internationales, nationales ou régionales. Les qualificateurs indiqués dans le tableau ci-après sont considérés comme ayant des valeurs de texte libre dépendant de la langue :

Tableau 5 : liste des qualificateurs avec des valeurs de texte libre dépendant de la langue pour les séquences de nucléotides

Section	Qualificateur de texte libre dépendant de la langue
6.3	bound_moiety
6.5	cell_type
6.8	clone
6.9	clone_lib
6.11	collected_by
6.14	cultivar
6.15	dev_stage
6.18	ecotype
6.21	frequency
6.22	function
6.24	gene_synonym
6.26	haplogroup
6.28	host
6.29	identified_by
6.30	isolate
6.31	isolation_source
6.32	lab_host
6.36	mating_type
6.41	note
6.45	organism
6.47	phenotype
6.49	pop_variant
6.50	product
6.66	serotype
6.67	serovar
6.68	sex
6.69	standard_name
6.70	strain
6.71	sub_clone
6.72	sub_species
6.73	sub_strain
6.75	tissue_lib
6.76	tissue_type
6.81	variety

6.1.	Qualifier	allele
	Definition	name of the allele for the given gene
	Mandatory value format	free text (NOTE: this value may require translation for National/Regional procedures)
	Example	<INSDQualifier_value>adh1-1</INSDQualifier_value>
	Comment	all gene-related features (exon, CDS etc) for a given gene should share the same allele qualifier value; the allele qualifier value must, by definition, be different from the gene qualifier value; when used with the variation feature key, the allele qualifier value should be that of the variant.

6.2.	Qualifier	anticodon
	Definition	location of the anticodon of tRNA and the amino acid for which it codes
	Mandatory value format	(pos:<location>,aa:<amino_acid>,seq:<text>) where <location> is the position of the anticodon and <amino_acid> is the three letter abbreviation for the amino acid encoded and <text> is the sequence of the anticodon
	Example	<INSDQualifier_value>(pos:34..36,aa:Phe,seq:aaa)</INSDQualifier_value> <INSDQualifier_value>(pos:join(5,495..496),aa:Leu,seq:taa)</INSDQualifier_value> <INSDQualifier_value>(pos:complement(4156..4158),aa:Glu,seq:ttg)</INSDQualifier_value>

6.3.	Qualifier	bound_moiety
	Definition	name of the molecule/complex that may bind to the given feature
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>GAL4</INSDQualifier_value>
	Comment	A single bound_moiety qualifier is permitted on the "misc_binding", "oriT" and "protein_bind" features.

6.4.	Qualifier	cell_line
	Definition	cell line from which the sequence was obtained
	Mandatory value format	free text
	Example	<INSDQualifier_value>MCF7</INSDQualifier_value>

6.5.	Qualifier	cell_type
	Definition	cell type from which the sequence was obtained
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>leukocyte</INSDQualifier_value>
6.6.	Qualifier	chromosome
	Definition	chromosome (e.g., Chromosome number) from which the sequence was obtained
	Mandatory value format	free text
	Example	<INSDQualifier_value>1</INSDQualifier_value>
6.7.	Qualifier	circular_RNA
	Definition	indicates that exons are out-of-order or overlapping because this spliced RNA product is a circular RNA (circRNA) created by backsplicing, for example when a downstream exon in the gene is located 5' of an upstream exon in the RNA product
	Value format	none
	Comment	should be used on features such as CDS, mRNA, tRNA and other features that are produced as a result of a backsplicing event. This qualifier should be used only when the splice event is indicated in the "join" operator, eg join(complement(69611..69724),139856..140087)
6.8.	Qualifier	clone
	Definition	clone from which the sequence was obtained
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>lambda-hIL7.3</INSDQualifier_value>
	Comment	a source feature must not contain more than one clone qualifier; where the sequence was obtained from multiple clones it may be further described in the feature table using the feature key misc_feature and a note qualifier to specify the multiple clones.
6.9.	Qualifier	clone_lib
	Definition	clone library from which the sequence was obtained
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>lambda-hIL7</INSDQualifier_value>

6.10.	Qualifier	codon_start
	Definition	indicates the offset at which the first complete codon of a coding feature can be found, relative to the first base of that feature.
	Mandatory value format	1 or 2 or 3
	Example	<INSDQualifier_value>2</INSDQualifier_value>

6.11.	Qualifier	collected_by
	Definition	name of persons or institute who collected the specimen
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>Dan Janzen</INSDQualifier_value>

6.12.	Qualifier	collection_date
	Definition	date that the specimen was collected.
	Mandatory value format	YYYY-MM-DD, YYYY-MM or YYYY
	Example	<INSDQualifier_value>1952-10-21</INSDQualifier_value> <INSDQualifier_value>1952-10</INSDQualifier_value> <INSDQualifier_value>1952</INSDQualifier_value>
	Comment	'YYYY' is a four-digit value representing the year. 'MM' is a two-digit value representing the month. 'DD' is a two-digit value representing the day of the month.

6.13.	Qualifier	compare
	Definition	Reference details of an existing public INSD entry to which a comparison is made
	Mandatory value format	[accession-number.sequence-version]
	Example	<INSDQualifier_value>AJ634337.1</INSDQualifier_value>
	Comment	This qualifier may be used on the following features: misc_difference, unsure, and variation. Multiple compare qualifiers with different contents are allowed within a single feature. This qualifier is not intended for large-scale annotation of variations, such as SNPs.

6.14.	Qualifier	cultivar
	Definition	cultivar (cultivated variety) of plant from which sequence was obtained
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>Nipponbare</INSDQualifier_value> <INSDQualifier_value>Tenuifolius</INSDQualifier_value> <INSDQualifier_value>Candy Cane</INSDQualifier_value> <INSDQualifier_value>IR36</INSDQualifier_value>
	Comment	'cultivar' is applied solely to products of artificial selection; use the variety qualifier for natural, named plant and fungal varieties.

6.15.	Qualifier	dev_stage
	Definition	if the sequence was obtained from an organism in a specific developmental stage, it is specified with this qualifier
	Mandatory value format	free textLanguage-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>fourth instar larva</INSDQualifier_value>

6.16.	Qualifier	direction
	Definition	direction of DNA replication
	Mandatory value format	left, right, or both where left indicates toward the 5' end of the sequence (as presented) and right indicates toward the 3' end
	Example	<INSDQualifier_value>left</INSDQualifier_value>
	Comment	The values left, right, and both are permitted when the direction qualifier is used to annotate a rep_origin feature key. However, only left and right values are permitted when the direction qualifier is used to annotate an oriT feature key.

6.17.	Qualifier	EC_number
	Definition	Enzyme Commission number for enzyme product of sequence
	Mandatory value format	free text
	Example	<INSDQualifier_value>1.1.2.4</INSDQualifier_value> <INSDQualifier_value>1.1.2.-</INSDQualifier_value> <INSDQualifier_value>1.1.2.n</INSDQualifier_value> <INSDQualifier_value>1.1.2.n1</INSDQualifier_value>
	Comment	valid values for EC numbers are defined in the list prepared by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) (published in Enzyme Nomenclature 1992, Academic Press, San Diego, or a more recent revision thereof).The format represents a string of four numbers separated by full stops; up to three numbers starting from the end of the string may be replaced by dash "-" to indicate uncertain assignment. Symbols including an "n", e.g., "n", "n1" and so on, may be used in the last position instead of a number where the EC number is awaiting assignment. Please note that such incomplete EC numbers are not approved by NC-IUBMB.

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

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

6.20.	Qualifier	exception
	Definition	indicates that the coding region cannot be translated using standard biological rules
	Mandatory value format	One of the following controlled vocabulary phrases: RNA editing rearrangement required for product annotated by transcript or proteomic data
	Example	<INSDQualifier_value>RNA editing</INSDQualifier_value> <INSDQualifier_value>rearrangement required for product</INSDQualifier_value>
	Comment	only to be used to describe biological mechanisms such as RNA editing; protein translation of a CDS with an exception qualifier will be different from the corresponding conceptual translation; must not be used where transl_except qualifier would be adequate, e.g., in case of stop codon completion use.

6.21.	Qualifier	frequency
	Definition	frequency of the occurrence of a feature Mandatory value format free text representing the proportion of a population carrying the feature expressed as a fraction Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>23/108</INSDQualifier_value> <INSDQualifier_value>1 in 12</INSDQualifier_value> <INSDQualifier_value>0.85</INSDQualifier_value>

6.22.	Qualifier	function
	Definition	function attributed to a sequence
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>essential for recognition of cofactor </INSDQualifier_value>
	Comment	The function qualifier is used when the gene name and/or product name do not convey the function attributable to a sequence.
6.23.	Qualifier	gene
	Definition	symbol of the gene corresponding to a sequence region
	Mandatory value format	free text
	Example	<INSDQualifier_value>ilvE</INSDQualifier_value>
	Comment	Use gene qualifier to provide the gene symbol; use standard_name qualifier to provide the full gene name.
6.24.	Qualifier	gene_synonym
	Definition	synonymous, replaced, obsolete or former gene symbol
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>Hox-3.3</INSDQualifier_value> in a feature where the gene qualifier value is Hoxc6
	Comment	used where it is helpful to indicate a gene symbol synonym; when the gene_synonym qualifier is used, a primary gene symbol must always be indicated in a gene qualifier
6.25.	Qualifier	germline
	Definition	the sequence presented has not undergone somatic rearrangement as part of an adaptive immune response; it is the unrearranged sequence that was inherited from the parental germline
	Value format	none
	Comment	germline qualifier must not be used to indicate that the source of the sequence is a gamete or germ cell; germline and rearranged qualifiers must not be used in the same source feature; germline and rearranged qualifiers must only be used for molecules that can undergo somatic rearrangements as part of an adaptive immune response; these are the T-cell receptor (TCR) and immunoglobulin loci in the jawed vertebrates, and the unrelated variable lymphocyte receptor (VLR) locus in the jawless fish (lampreys and hagfish); germline and rearranged qualifiers should not be used outside of the Craniata (taxid=89593)

6.26.	Qualifier	haplogroup
	Definition	name for a group of similar haplotypes that share some sequence variation. Haplogroups are often used to track migration of population groups.
	Mandatory value format	free textLanguage-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>H*</INSDQualifier_value>

6.27.	Qualifier	haplotype
	Definition	name for a specific set of alleles that are linked together on the same physical chromosome. In the absence of recombination, each haplotype is inherited as a unit, and may be used to track gene flow in populations.
	Mandatory value format	free text
	Example	<INSDQualifier_value>Dw3 B5 Cw1 A1</INSDQualifier_value>

6.28.	Qualifier	host
	Definition	natural (as opposed to laboratory) host to the organism from which sequenced molecule was obtained
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>Homo sapiens</INSDQualifier_value> <INSDQualifier_value>Homo sapiens 12 year old girl</INSDQualifier_value> <INSDQualifier_value>Rhizobium NGR234</INSDQualifier_value>

6.29.	Qualifier	identified_by
	Definition	name of the expert who identified the specimen taxonomically
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>John Burns</INSDQualifier_value>

6.30.	Qualifier	isolate
	Definition	individual isolate from which the sequence was obtained
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>Patient #152</INSDQualifier_value> <INSDQualifier_value>DGGE band PSBAC-13</INSDQualifier_value>

6.31.	Qualifier	isolation_source
	Definition	describes the physical, environmental and/or local geographical source of the biological sample from which the sequence was derived
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Examples	<INSDQualifier_value>rumen isolates from standard Pelleted ration-fed steer #67</INSDQualifier_value> <INSDQualifier_value>permanent Antarctic sea ice</INSDQualifier_value> <INSDQualifier_value>denitrifying activated sludge from carbon_limited continuous reactor</INSDQualifier_value>
	Comment	used only with the source feature key; source feature keys containing an environmental_sample qualifier should also contain an isolation_source qualifier
6.32.	Qualifier	lab_host
	Definition	scientific name of the laboratory host used to propagate the source organism from which the sequenced molecule was obtained
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>Gallus gallus</INSDQualifier_value> <INSDQualifier_value>Gallus gallus embryo</INSDQualifier_value> <INSDQualifier_value>Escherichia coli strain DH5 alpha</INSDQualifier_value> <INSDQualifier_value>Homo sapiens HeLa cells</INSDQualifier_value>
	Comment	the full binomial scientific name of the host organism should be used when known; extra conditional information relating to the host may also be included
6.33.	Qualifier	lat_lon
	Definition	geographical coordinates of the location where the specimen was collected
	Mandatory value format	free text – degrees latitude and longitude in format "d[d.dddd] N|S d[dd.dddd] W|E"
	Example	<INSDQualifier_value>47.94 N 28.12 W</INSDQualifier_value> <INSDQualifier_value>45.0123 S 4.1234 E</INSDQualifier_value>
6.34.	Qualifier	macronuclear
	Definition	if the sequence shown is DNA and from an organism which undergoes chromosomal differentiation between macronuclear and micronuclear stages, this qualifier is used to denote that the sequence is from macronuclear DNA
	Value format	none

6.35.	Qualifier	map
	Definition	genomic map position of feature
	Mandatory value format	free text
	Example	<INSDQualifier_value>8q12-q13</INSDQualifier_value>

6.36.	Qualifier	mating_type
	Definition	mating type of the organism from which the sequence was obtained; mating type is used for prokaryotes, and for eukaryotes that undergo meiosis without sexually dimorphic gametes
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Examples	<INSDQualifier_value>MAT-1</INSDQualifier_value> <INSDQualifier_value>plus</INSDQualifier_value> <INSDQualifier_value>-</INSDQualifier_value> <INSDQualifier_value>odd</INSDQualifier_value> <INSDQualifier_value>even</INSDQualifier_value>“
	Comment	mating_type qualifier values male and female are valid in the prokaryotes, but not in the eukaryotes; for more information, see the entry for the sex qualifier.

6.37.	Qualifier	mobile_element_type
	Definition	type and name or identifier of the mobile element which is described by the parent feature
	Mandatory value format	<mobile_element_type>[:<mobile_element_name>] where <mobile_element_type> is one of the following: transposon retrotransposon integron insertion sequence non-LTR retrotransposon SINE MITE LINE other
	Example	<INSDQualifier_value>transposon:Tnp9</INSDQualifier_value>
	Comment	mobile_element_type is permitted on mobile_element feature key only. Mobile element should be used to represent both elements which are currently mobile, and those which were mobile in the past. Value “other” for <mobile_element_type> requires a <mobile_element_name>

6.38.	Qualifier	mod_base
	Definition	abbreviation for a modified nucleotide base
	Mandatory value format	modified base abbreviation chosen from this Annex, Section 2
	Example	<INSDQualifier_value>m5c</INSDQualifier_value> <INSDQualifier_value>OTHER</INSDQualifier_value>
	Comment	specific modified nucleotides not found in Section 2 of this Annex are annotated by entering OTHER as the value for the mod_base qualifier and including a note qualifier with the full name of the modified base as its value

6.39.	Qualifier	mol_type
	Definition	molecule type of sequence
	Mandatory value format	One chosen from the following: genomic DNA genomic RNA mRNA tRNA rRNA other RNA other DNA transcribed RNA viral cRNA unassigned DNA unassigned RNA
	Example	<INSDQualifier_value>genomic DNA</INSDQualifier_value> <INSDQualifier_value>other RNA</INSDQualifier_value>
	Comment	mol_type qualifier is mandatory on the source feature key; the value "genomic DNA" does not imply that the molecule is nuclear (e.g., organelle and plasmid DNA must be described using "genomic DNA"); ribosomal RNA genes must be described using "genomic DNA"; "rRNA" must only be used if the ribosomal RNA molecule itself has been sequenced; values "other RNA" and "other DNA" must be applied to synthetic molecules, values "unassigned DNA", "unassigned RNA" must be applied where in vivo molecule is unknown.

6.40.	Qualifier	ncRNA_class
	Definition	a structured description of the classification of the non-coding RNA described by the ncRNA parent key
	Mandatory value format	TYPE where TYPE is one of the following controlled vocabulary terms or phrases: antisense_RNA autocatalytically_spliced_intron circRNA ribozyme hammerhead_ribozyme lncRNA RNase_P_RNA RNase_MRP_RNA telomerase_RNA guide_RNA sgRNA rasiRNA scRNA scaRNA siRNA pre_miRNA

		miRNA piRNA snoRNA snRNA SRP_RNA vault_RNA Y_RNA other
	Example	<INSDQualifier_value>autocatalytically_spliced_intron </INSDQualifier_value> <INSDQualifier_value>siRNA</INSDQualifier_value> <INSDQualifier_value>scRNA</INSDQualifier_value> <INSDQualifier_value>other</INSDQualifier_value>
	Comment	specific ncRNA types not yet in the ncRNA_class controlled vocabulary must be annotated by entering "other" as the ncRNA_class qualifier value, and providing a brief explanation of novel ncRNA_class in a note qualifier
6.41.	Qualifier	note
	Definition	any comment or additional information
	Mandatory value format	free textLanguage-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>A comment about the feature</INSDQualifier_value>
6.42.	Qualifier	number
	Definition	a number to indicate the order of genetic elements (e.g., exons or introns) in the 5' to 3' direction
	Mandatory value format	free text (with no whitespace characters)
	Example	<INSDQualifier_value>4</INSDQualifier_value> <INSDQualifier_value>6B</INSDQualifier_value>
	Comment	text limited to integers, letters or combination of integers and/or letters represented as a data value that contains no whitespace characters; any additional terms should be included in a standard_name qualifier. Example: a number qualifier with a value of 2A and a standard_name qualifier with a value of "long"
6.43.	Qualifier	operon
	Definition	name of the group of contiguous genes transcribed into a single transcript to which that feature belongs
	Mandatory value format	free text
	Example	<INSDQualifier_value>lac</INSDQualifier_value>
6.44.	Qualifier	organelle
	Definition	type of membrane-bound intracellular structure from which the sequence was obtained
	Mandatory value format	One of the following controlled vocabulary terms and phrases: chromatophore hydrogenosome mitochondrion nucleomorph plastid mitochondrion:kinetoplast

		plastid:chloroplast plastid:apicoplast plastid:chromoplast plastid:cyanelle plastid:leucoplast plastid:proplastid
	Examples	<INSDQualifier_value>chromatophore</INSDQualifier_value> <INSDQualifier_value>hydrogenosome</INSDQualifier_value> <INSDQualifier_value>mitochondrion</INSDQualifier_value> <INSDQualifier_value>nucleomorph</INSDQualifier_value> <INSDQualifier_value>plastid</INSDQualifier_value> <INSDQualifier_value>mitochondrion:kinetoplast</INSDQualifier_value> <INSDQualifier_value>plastid:chloroplast</INSDQualifier_value> <INSDQualifier_value>plastid:apicoplast</INSDQualifier_value> <INSDQualifier_value>plastid:chromoplast</INSDQualifier_value> <INSDQualifier_value>plastid:cyanelle</INSDQualifier_value> <INSDQualifier_value>plastid:leucoplast</INSDQualifier_value> <INSDQualifier_value>plastid:proplastid</INSDQualifier_value>
6.45.	Qualifier	organism
	Definition	scientific name of the organism that provided the sequenced genetic material, if known, or the available taxonomic information if the organism is unclassified; or an indication that the sequence is a synthetic construct
	Mandatory value format	free textLanguage-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>Homo sapiens</INSDQualifier_value>
6.46.	Qualifier	PCR_primers
	Definition	PCR primers that were used to amplify the sequence. A single PCR_primers qualifier should contain all the primers used for a single PCR reaction. If multiple forward or reverse primers are present in a single PCR reaction, multiple sets of fwd_name/fwd_seq or rev_name/rev_seq values will be present
	Mandatory value format	[fwd_name: XXX1,]fwd_seq: xxxxx1,[fwd_name: XXX2,]fwd_seq: xxxxx2, [rev_name: YYY1,]rev_seq: yyyyy1,[rev_name: YYY2,]rev_seq: yyyyy2
	Example	<INSDQualifier_value>fwd_name: CO1P1, fwd_seq: ttgatttttggtcayccwgaagt,rev_name: CO1R4, rev_seq: ccwvytardcctarraartgttg</INSDQualifier_value> <INSDQualifier_value>fwd_name: hoge1, fwd_seq: cgkgtgtatcttact, rev_name: hoge2, rev_seq: cg<i>gtgtatcttact</INSDQualifier_value> <INSDQualifier_value>fwd_name: CO1P1, fwd_seq: ttgatttttggtcayccwgaagt, fwd_name: CO1P2, fwd_seq: gatacacaggtcayccwgaagt, rev_name: CO1R4, rev_seq: ccwvytardcctarraartgttg</INSDQualifier_value>
	Comment	fwd_seq and rev_seq are both mandatory; fwd_name and rev_name are both optional. Both sequences must be presented in 5' to 3' order. The sequences must be given in the symbols from Section 1 of this Annex, except for the modified bases, which must be enclosed within angle brackets < >. In XML, the angle brackets < and > must be substituted with < and > since they are reserved characters in XML.
6.47.	Qualifier	phenotype
	Definition	phenotype conferred by the feature, where phenotype is defined as a physical, biochemical or behavioural characteristic or set of characteristics
	Mandatory value format	free textLanguage-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>erythromycin resistance</INSDQualifier_value>

6.48.	Qualifier	plasmid
	Definition	name of naturally occurring plasmid from which the sequence was obtained, where plasmid is defined as an independently replicating genetic unit that cannot be described by chromosome or segment qualifiers
	Mandatory value format	free text
	Example	<INSDQualifier_value>pC589</INSDQualifier_value>
6.49.	Qualifier	pop_variant
	Definition	name of subpopulation or phenotype of the sample from which the sequence was derived
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>pop1</INSDQualifier_value> <INSDQualifier_value>Bear Paw</INSDQualifier_value>
6.50.	Qualifier	product
	Definition	name of the product associated with the feature, e.g., the mRNA of an mRNA feature, the polypeptide of a CDS, the mature peptide of a mat_peptide, etc.
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>trypsinogen</INSDQualifier_value> (when qualifier appears in CDS feature) <INSDQualifier_value>trypsin</INSDQualifier_value> (when qualifier appears in mat_peptide feature) <INSDQualifier_value>XYZ neural-specific transcript</INSDQualifier_value> (when qualifier appears in mRNA feature)
6.51.	Qualifier	protein_id
	Definition	protein sequence identification number, an integer used in a sequence listing to designate the protein sequence encoded by the coding sequence identified in the corresponding CDS feature key and translation qualifier
	Mandatory value format	an integer greater than zero
	Example	<INSDQualifier_value>89</INSDQualifier_value>
6.52.	Qualifier	proviral
	Definition	this qualifier is used to flag sequence obtained from a virus or phage that is integrated into the genome of another organism
	Value format	none

6.53.	Qualifier	pseudo
	Definition	indicates that this feature is a non-functional version of the element named by the feature key
	Value format	none
	Comment	The qualifier pseudo should be used to describe non-functional genes that are not formally described as pseudogenes, e.g., CDS has no translation due to other reasons than pseudogenization events. Other reasons may include sequencing or assembly errors. In order to annotate pseudogenes the qualifier pseudogene must be used, indicating the TYPE of pseudogene.
6.54.	Qualifier	pseudogene
	Definition	indicates that this feature is a pseudogene of the element named by the feature key
	Mandatory value format	TYPE where TYPE is one of the following controlled vocabulary terms or phrases: processed unprocessed unitary allelic unknown
	Example	<INSDQualifier_value>processed</INSDQualifier_value> <INSDQualifier_value>unprocessed</INSDQualifier_value> <INSDQualifier_value>unitary</INSDQualifier_value> <INSDQualifier_value>allelic</INSDQualifier_value> <INSDQualifier_value>unknown</INSDQualifier_value>
	Comment	Definitions of TYPE values: processed – the pseudogene has arisen by reverse transcription of a mRNA into cDNA, followed by reintegration into the genome. Therefore, it has lost any intron/exon structure, and it might have a pseudo-polyA-tail. unprocessed – the pseudogene has arisen from a copy of the parent gene by duplication followed by accumulation of random mutations. The changes, compared to their functional homolog, include insertions, deletions, premature stop codons, frameshifts and a higher proportion of non-synonymous versus synonymous substitutions. unitary – the pseudogene has no parent. It is the original gene, which is functional is some species but disrupted in some way (indels, mutation, recombination) in another species or strain. allelic – a (unitary) pseudogene that is stable in the population but importantly it has a functional alternative allele also in the population. i.e., one strain may have the gene, another strain may have the pseudogene. MHC haplotypes have allelic pseudogenes. unknown – the submitter does not know the method of pseudogenization.
6.55.	Qualifier	rearranged
	Definition	the sequence presented in the entry has undergone somatic rearrangement as part of an adaptive immune response; it is not the unrearranged sequence that was inherited from the parental germline
	Value format	none
	Comment	The rearranged qualifier must not be used to annotate chromosome rearrangements that are not involved in an adaptive immune response; germline and rearranged qualifiers must not be used in the same source feature; germline and rearranged qualifiers must only be used for molecules that can undergo somatic rearrangements as part of an adaptive immune response; these are the T-cell receptor (TCR) and immunoglobulin loci in the jawed vertebrates, and the unrelated variable lymphocyte receptor (VLR) locus in the jawless fish (lampreys and hagfish); germline and rearranged qualifiers should not be used outside of the Craniata (taxid=89593)

6.56.	Qualifier	recombination_class
	Definition	a structured description of the classification of recombination hotspot region within a sequence
	Mandatory value format	TYPE where TYPE is one of the following controlled vocabulary terms or phrases: meiotic mitotic non_allelic_homologous chromosome_breakpoint other
	Example	<INSDQualifier_value>meiotic </INSDQualifier_value> <INSDQualifier_value>chromosome_breakpoint</INSDQualifier_value>
	Comment	specific recombination classes not yet in the recombination_class controlled vocabulary must be annotated by entering "other" as the recombination_class qualifier value and providing a brief explanation of the novel recombination_class in a note qualifier
6.57.	Qualifier	regulatory_class
	Definition	a structured description of the classification of transcriptional, translational, replicational and chromatin structure related regulatory elements in a sequence
	Mandatory value format	TYPE where TYPE is one of the following controlled vocabulary terms or phrases: attenuator CAAT_signal DNase_I_hypersensitive_site enhancer enhancer_blocking_element GC_signal imprinting_control_region insulator locus_control_region matrix_attachment_region minus_35_signal minus_10_signal polyA_signal_sequence promoter recoding_stimulatory_region recombination_enhancerreplication_regulatory_region response_element promoter ribosome_binding_site riboswitch silencer TATA_box terminator transcriptional_cis_regulatory_region uORF other
	Example	<INSDQualifier_value>promoter</INSDQualifier_value> <INSDQualifier_value>enhancer</INSDQualifier_value> <INSDQualifier_value>ribosome_binding_site</INSDQualifier_value>
	Comment	specific regulatory classes not yet in the regulatory_class controlled vocabulary must be annotated by entering "other" as the regulatory_class qualifier value and providing a brief explanation of the novel regulatory_class in a note qualifier

6.58.	Qualifier	replace
	Definition	indicates that the sequence identified in a feature's location is replaced by the sequence shown in the qualifier's value; if no sequence (i.e., no value) is contained within the qualifier, this indicates a deletion
	Mandatory value format	free text
	Example	<INSDQualifier_value>a</INSDQualifier_value> <INSDQualifier_value></INSDQualifier_value> – for a deletion
6.59.	Qualifier	ribosomal_slippage
	Definition	during protein translation, certain sequences can program ribosomes to change to an alternative reading frame by a mechanism known as ribosomal slippage
	Value format	none
	Comment	a join operator, e.g.,: [join(486..1784,1787..4810)] must be used in the CDS feature location to indicate the location of ribosomal_slippage
6.60.	Qualifier	rpt_family
	Definition	type of repeated sequence; "Alu" or "Kpn", for example
	Mandatory value format	free text
	Example	<INSDQualifier_value>Alu</INSDQualifier_value>
6.61.	Qualifier	rpt_type
	Definition	structure and distribution of repeated sequence
	Mandatory value format	One of the following controlled vocabulary terms or phrases: tandem direct inverted flanking nested terminal dispersed long_terminal_repeat non_ltr_retrotransposon_polymeric_tract centromeric_repeat telomeric_repeat x_element_combinatorial_repeat y_prime_element other
	Example	<INSDQualifier_value>inverted</INSDQualifier_value> <INSDQualifier_value>long_terminal_repeat</INSDQualifier_value>
	Comment	Definitions of the values: tandem – a repeat that exists adjacent to another in the same orientation; direct – a repeat that exists not always adjacent but is in the same orientation; inverted – a repeat pair occurring in reverse orientation to one another on the same molecule; flanking – a repeat lying outside the sequence for which it has functional significance (eg. transposon insertion target sites); nested – a repeat that is disrupted by the insertion of another element; dispersed – a repeat that is found dispersed throughout the genome; terminal – a repeat at the ends of and within the sequence for which it has functional significance (eg. transposon LTRs); long_terminal_repeat – a sequence directly repeated at both ends of a defined sequence, of the sort typically found in retroviruses;

		non_ltr_retrotransposon_polymeric_tract – a polymeric tract, such as poly(dA), within a non LTR retrotransposon; centromeric_repeat – a repeat region found within the modular centromere; telomeric_repeat – a repeat region found within the telomere; x_element_combinatorial_repeat – a repeat region located between the X element and the telomere or adjacent Y' element; y_prime_element – a repeat region located adjacent to telomeric repeats or X element combinatorial repeats, either as a single copy or tandem repeat of two to four copies; other – a repeat exhibiting important attributes that cannot be described by other values.
6.62.	Qualifier	rpt_unit_range
	Definition	location of a repeating unit expressed as a range
	Mandatory value format	<base_range> – where <base_range> is the first and last base (separated by two dots) of a repeating unit
	Example	<INSDQualifier_value>202..245</INSDQualifier_value>
	Comment	used to indicate the base range of the sequence that constitutes a repeating unit within the region specified by the feature keys oriT and repeat_region.
6.63.	Qualifier	rpt_unit_seq
	Definition	identity of a repeat sequence
	Mandatory value format	free text
	Example	<INSDQualifier_value>aagggc</INSDQualifier_value> <INSDQualifier_value>ag(5)tg(8)</INSDQualifier_value> <INSDQualifier_value>(AAAGA)6(AAAA)1(AAAGA)12</INSDQualifier_value>
	Comment	used to indicate the literal sequence that constitutes a repeating unit within the region specified by the feature keys oriT and repeat_region
6.64.	Qualifier	satellite
	Definition	identifier for a satellite DNA marker, composed of many tandem repeats (identical or related) of a short basic repeated unit
	Mandatory value format	<satellite_type>[:<class>][<identifier>] – where <satellite_type> is one of the following: satellite microsatellite minisatellite
	Example	<INSDQualifier_value>satellite: S1a</INSDQualifier_value> <INSDQualifier_value>satellite: alpha</INSDQualifier_value> <INSDQualifier_value>satellite: gamma III</INSDQualifier_value> <INSDQualifier_value>microsatellite: DC130</INSDQualifier_value>
	Comment	many satellites have base composition or other properties that differ from those of the rest of the genome that allows them to be identified.
6.65.	Qualifier	segment
	Definition	name of viral or phage segment sequenced
	Mandatory value format	free text
	Example	<INSDQualifier_value>6</INSDQualifier_value>

6.66.	Qualifier	serotype
	Definition	serological variety of a species characterized by its antigenic properties
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>B1</INSDQualifier_value>
	Comment	used only with the source feature key; the Bacteriological Code recommends the use of the term 'serovar' instead of 'serotype' for the prokaryotes; see the International Code of Nomenclature of Bacteria (1990 Revision) Appendix 10.B "Infraspecific Terms".
6.67.	Qualifier	serovar
	Definition	serological variety of a species (usually a prokaryote) characterized by its antigenic properties
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>O157:H7</INSDQualifier_value>
	Comment	used only with the source feature key; the Bacteriological Code recommends the use of the term 'serovar' instead of 'serotype' for prokaryotes; see the International Code of Nomenclature of Bacteria (1990 Revision) Appendix 10.B "Infraspecific Terms".
6.68.	Qualifier	sex
	Definition	sex of the organism from which the sequence was obtained; sex is used for eukaryotic organisms that undergo meiosis and have sexually dimorphic gametes
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Examples	<INSDQualifier_value>female</INSDQualifier_value> <INSDQualifier_value>male</INSDQualifier_value> <INSDQualifier_value>hermaphrodite</INSDQualifier_value> <INSDQualifier_value>unisexual</INSDQualifier_value> <INSDQualifier_value>bisexual</INSDQualifier_value> <INSDQualifier_value>asexual</INSDQualifier_value> <INSDQualifier_value>monoecious</INSDQualifier_value> [or monecious] <INSDQualifier_value>dioecious</INSDQualifier_value> [or diecious]
	Comment	The sex qualifier should be used (instead of mating_type qualifier) in the Metazoa, Embryophyta, Rhodophyta & Phaeophyceae; mating_type qualifier should be used (instead of sex qualifier) in the Bacteria, Archaea & Fungi; neither sex nor mating_type qualifiers should be used in the viruses; outside of the taxa listed above, mating_type qualifier should be used unless the value of the qualifier is taken from the vocabulary given in the examples above
6.69.	Qualifier	standard_name
	Definition	accepted standard name for this feature
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>dotted</INSDQualifier_value>

	Comment	use standard_name qualifier to give full gene name, but use gene qualifier to give gene symbol (in the above example gene qualifier value is Dt).

6.70.	Qualifier	strain
	Definition	strain from which sequence was obtained
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>BALB/c</INSDQualifier_value>
	Comment	feature entries including a strain qualifier must not include the environmental_sample qualifier

6.71.	Qualifier	sub_clone
	Definition	sub-clone from which sequence was obtained
	Mandatory value format	free textLanguage-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>lambda-hIL7.20g</INSDQualifier_value>
	Comment	a source feature must not contain more than one sub_clone qualifier; to indicate that the sequence was obtained from multiple sub_clones, multiple sources may be further described using the feature key "misc_feature" and the qualifier "note"

6.72.	Qualifier	sub_species
	Definition	name of sub-species of organism from which sequence was obtained
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>lactis</INSDQualifier_value>

6.73.	Qualifier	sub_strain
	Definition	name or identifier of a genetically or otherwise modified strain from which sequence was obtained, derived from a parental strain (which should be annotated in the strain qualifier). sub_strain from which sequence was obtained
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>abis</INSDQualifier_value>
	Comment	must be accompanied by a strain qualifier in a source feature; if the parental strain is not given, the modified strain should be annotated in the strain qualifier instead of sub_strain. For example, either a strain qualifier with the value K-12 and a substrain qualifier with the value MG1655 or a strain qualifier with the value MG1655

6.74.	Qualifier	tag_peptide
	Definition	base location encoding the polypeptide for proteolysis tag of tmRNA and its termination codon
	Mandatory value format	<base_range> – where <base_range> provides the first and last base (separated by two dots) of the location for the proteolysis tag
	Example	<INSDQualifier_value>90..122</INSDQualifier_value>
	Comment	it is recommended that the amino acid sequence corresponding to the tag_peptide be annotated by describing a 5' partial CDS feature; e.g., CDS with a location of <90..122
6.75.	Qualifier	tissue_lib
	Definition	tissue library from which sequence was obtained
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>tissue library 772</INSDQualifier_value>
6.76.	Qualifier	tissue_type
	Definition	tissue type from which the sequence was obtained
	Mandatory value format	free textLanguage-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>liver</INSDQualifier_value>
6.77.	Qualifier	transl_except
	Definition	translational exception: single codon the translation of which does not conform to genetic code defined by organism or transl_table.
	Mandatory value format	(pos:<location>aa:<amino_acid>) where <amino_acid> is the three letter abbreviation for the amino acid coded by the codon at the base_range position
	Example	<INSDQualifier_value>(pos:213..215,aa:Trp)<//INSDQualifier_value> <INSDQualifier_value>(pos:462..464,aa:OTHER)<//INSDQualifier_value> <INSDQualifier_value>(pos:1017,aa:TERM)<//INSDQualifier_value> <INSDQualifier_value>(pos:2000..2001,aa:TERM)<//INSDQualifier_value>
	Comment	if the amino acid is not one of the specific amino acids listed in Section 3 of this Annex, use OTHER as <amino_acid> and provide the name of the unusual amino acid in a note qualifier; for modified amino-acid selenocysteine use three letter abbreviation 'Sec' (one letter symbol 'U' in amino-acid sequence) for <amino _acid>; for modified amino-acid pyrrolysine use three letter abbreviation 'Pyl' (one letter symbol 'O' in amino-acid sequence) for <amino _acid>; for partial termination codons where TAA stop codon is completed by the addition of 3' A residues to the mRNA either a single base_position or a base_range is used for the location, see the third and fourth examples above, in conjunction with a note qualifier indicating 'stop codon completed by the addition of 3' A residues to the mRNA'.

6.78.	Qualifier	transl_table
	Definition	definition of genetic code table used if other than universal or standard genetic code table. Tables used are described in this Annex
	Mandatory value format	<integer> where <integer> is the number assigned to the genetic code table
	Example	<INSDQualifier_value>3</INSDQualifier_value> – example where the yeast mitochondrial code is to be used
	Comment	if the transl_table qualifier is not used to further annotate a CDS feature key, then the CDS is translated using the Standard Code (i.e. Universal Genetic Code). Genetic code exceptions outside the range of specified tables are reported in transl_except qualifiers.

6.79.	Qualifier	trans_splicing
	Definition	indicates that exons from two RNA molecules are ligated in intermolecular reaction to form mature RNA
	Value format	none
	Comment	should be used on features such as CDS, mRNA and other features that are produced as a result of a trans-splicing event. This qualifier must be used only when the splice event is indicated in the "join" operator, e.g., join (complement(69611..69724),139856..140087) in the feature location

6.80.	Qualifier	translation
	Definition	one-letter abbreviated amino acid sequence derived from either the standard (or universal) genetic code or the table as specified in a transl_table qualifier and as determined by an exception in the transl_except qualifier
	Mandatory value format	contiguous string of one-letter amino acid abbreviations from Section 3 of this Annex, "X" is to be used for AA exceptions.
	Example	<INSDQualifier_value>MASTFPPWYRGCASTPSLKGLIMCTW</INSDQualifier_value>
	Comment	to be used with CDS feature only; must be accompanied by protein_id qualifier when the translation product contains four or more specifically defined amino acids; see transl_table for definition and location of genetic code Tables; only one of the qualifiers translation, pseudo and pseudogene are permitted to further annotate a CDS feature.

6.81.	Qualifier	variety
	Definition	variety (= varietas, a formal Linnaean rank) of organism from which sequence was derived.
	Mandatory value format	free textLanguage-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>insularis</INSDQualifier_value>
	Comment	use the cultivar qualifier for cultivated plant varieties, i.e., products of artificial selection; varieties other than plant and fungal variatas should be annotated via a note qualifier, e.g., with the value <INSDQualifier_value>breed:Cukorova</INSDQualifier_value>

SECTION 7 : CLÉS DE CARACTÉRISATION POUR LES SÉQUENCES D'ACIDES AMINÉS

La présente section contient la liste des clés de caractérisation pouvant être employées pour les séquences d'acides aminés. Les clés de caractérisation sont présentées dans l'ordre alphabétique.

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

7.7.	Feature Key	COMPBIAS
	Definition	Extent of a compositionally biased region
	Optional qualifiers	note

7.8.	Feature Key	CONFLICT
	Definition	Different sources report differing sequences
	Optional qualifiers	note
	Comment	Examples of values for the "note" qualifier: Missing; K -> Q; GSDSE -> RIRLR; V -> A.

7.9.	Feature Key	CROSSLNK
	Definition	Post translationally formed amino acid bonds
	Mandatory qualifiers	note
	Comment	Covalent linkages of various types formed between two proteins (interchain cross-links) or between two parts of the same protein (intrachain cross-links); except for cross-links formed by disulfide bonds, for which the "DISULFID" feature key is to be used. For an interchain cross-link, the location descriptor in the feature location element is the residue number of the amino acid cross-linked to the other protein. For an intrachain cross-link, the location descriptor in the feature location element is the residue numbers of the cross-linked amino acids in"x..y" format, e.g. "42.. 50". The note qualifier indicates the nature of the cross-link; at least specifying the name of the conjugate and the identity of the two amino acids involved. Examples of values for the "note" qualifier: "Isoglutamyl cysteine thioester (Cys-Gln);" "Beta-methyllanthionine (Cys-Thr);" and "Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in ubiquitin)"

7.10.	Feature Key	DISULFID
	Definition	Disulfide bond
	Mandatory	note
	Comment	For an interchain disulfide bond, the location descriptor in the feature location element is the residue number of the cysteine linked to the other protein. For an intrachain cross-link, the location descriptor in the feature location element is the residue numbers of the linked cysteines in"x y" format, e.g. "42.. 50". For interchain disulfide bonds, the note qualifier indicates the nature of the cross-link, by identifying the other protein, for example, "Interchain (between A and B chains)"

7.11.	Feature Key	DNA_BIND
	Definition	Extent of a DNA-binding region
	Mandatory qualifiers	note
	Comment	The nature of the DNA-binding region is given in the note qualifier. Examples of values for the "note" qualifier: "Homeobox" and "Myb 2"

7.12.	Feature Key	DOMAIN
	Definition	Extent of a domain, which is defined as a specific combination of secondary structures organized into a characteristic three-dimensional structure or fold
	Mandatory qualifiers	note
	Comment	The domain type is given in the note qualifier. Where several copies of a domain are present, the domains are numbered. Examples of values for the "note" qualifier: "Ras-GAP" and "Cadherin 1"

7.13.	Feature Key	HELIX
	Definition	Secondary structure: Helices, for example, Alpha-helix; 3(10) helix; or Pi-helix
	Optional qualifiers	note
	Comment	This feature is used only for proteins whose tertiary structure is known. Only three types of secondary structure are specified: helices (key HELIX), beta-strands (key STRAND) and turns (key TURN). Residues not specified in one of these classes are in a 'loop' or 'random-coil' structure.

7.14.	Feature Key	INIT_MET
	Definition	Initiator methionine
	Optional qualifiers	note
	Comment	The location descriptor in the feature location element is "1". This feature key indicates the N-terminal methionine is cleaved off. This feature is not used when the initiator methionine is not cleaved off.

7.15.	Feature Key	INTRAMEM
	Definition	Extent of a region located in a membrane without crossing it
	Optional qualifiers	note

7.16.	Feature Key	LIPID
	Definition	Covalent binding of a lipid moiety
	Mandatory qualifiers	note
	Comment	The chemical nature of the bound lipid moiety is given in the note qualifier, indicating at least the name of the lipidated amino acid. Examples of values for the "note" qualifier: "N-myristoyl glycine"; "GPI-anchor amidated serine" and "S-diacylglycerol cysteine."

7.17.	Feature Key	METAL
	Definition	Binding site for a metal ion.
	Mandatory qualifiers	note
	Comment	The note qualifier indicates the nature of the metal. Examples of values for the "note" qualifier: "Iron (heme axial ligand)" and "Copper".

7.18.	Feature Key	MOD_RES
	Definition	Posttranslational modification of a residue
	Mandatory qualifiers	note
	Comment	The chemical nature of the modified residue is given in the note qualifier, indicating at least the name of the post-translationally modified amino acid. If the modified amino acid is listed in Section 4 of this Annex, the abbreviation may be used in place of the the full name. Examples of values for the "note" qualifier: "N-acetylalanine"; "3-Hyp"; and "MeLys" or "N-6-methyllysine"
7.19.	Feature Key	MOTIF
	Definition	Short (up to 20 amino acids) sequence motif of biological interest
	Optional qualifiers	note
7.20.	Feature Key	MUTAGEN
	Definition	Site which has been experimentally altered by mutagenesis
	Optional qualifiers	note
7.21.	Feature Key	NON_STD
	Definition	Non-standard amino acid
	Optional qualifiers	note
	Comment	This key only describes the occurrence of non-standard amino acids selenocysteine (U) and pyrrolysine (O) in the amino acid sequence.
7.22.	Feature Key	NON_TER
	Definition	The residue at an extremity of the sequence is not the terminal residue
	Optional qualifiers	note
	Comment	If applied to position 1, this means that the first position is not the N-terminus of the complete molecule. If applied to the last position, it means that this position is not the C-terminus of the complete molecule.
7.23.	Feature Key	NP_BIND
	Definition	Extent of a nucleotide phosphate-binding region
	Mandatory qualifiers	note
	Comment	The nature of the nucleotide phosphate is indicated in the note qualifier. Examples of values for the "note" qualifier: "ATP" and "FAD".
7.24.	Feature Key	PEPTIDE
	Definition	Extent of a released active peptide
	Optional qualifiers	note

7.25.	Feature Key	PROPEP
	Definition	Extent of a propeptide
	Optional qualifiers	note

7.26.	Feature Key	REGION
	Definition	Extent of a region of interest in the sequence
	Optional qualifiers	note

7.27.	Feature Key	REPEAT
	Definition	Extent of an internal sequence repetition
	Optional qualifiers	note

7.28.	Feature Key	SIGNAL
	Definition	Extent of a signal sequence (prepeptide)
	Optional qualifiers	note

7.29.	Feature Key	SITE
	Definition	Any interesting single amino-acid site on the sequence that is not defined by another feature key. It can also apply to an amino acid bond which is represented by the positions of the two flanking amino acids
	Mandatory qualifier	note
	Comment	When SITE is used to annotate a modified amino acid the value for the qualifier "note" must either be an abbreviation set forth in Section 4 of this Annex, or the complete, unabbreviated name of the modified amino acid.

7.30.	Feature Key	source
	Definition	Identifies the source of the sequence; this key is mandatory; every sequence will have a single source feature spanning the entire sequence
	Mandatory qualifiers	mol_type organism
	Optional qualifiers	note

7.31.	Feature Key	STRAND
	Definition	Secondary structure: Beta-strand; for example Hydrogen bonded beta-strand or residue in an isolated beta-bridge
	Optional qualifiers	note
	Comment	This feature is used only for proteins whose tertiary structure is known. Only three types of secondary structure are specified: helices (key HELIX), beta-strands (key STRAND) and turns (key TURN). Residues not specified in one of these classes are in a 'loop' or 'random-coil' structure.

7.32.	Feature Key	TOPO_DOM
	Definition	Topological domain
	Optional qualifiers	note

7.33.	Feature Key	TRANSMEM
	Definition	Extent of a transmembrane region
	Optional qualifiers	note

7.34.	Feature Key	TRANSIT
	Definition	Extent of a transit peptide (mitochondrion, chloroplast, thylakoid, cyanelle, peroxisome etc.)
	Optional qualifiers	note

7.35.	Feature Key	TURN
	Definition	Secondary structure Turns, for example, H-bonded turn (3-turn, 4-turn or 5-turn)
	Optional qualifiers	note
	Comment	This feature is used only for proteins whose tertiary structure is known. Only three types of secondary structure are specified: helices (key HELIX), beta-strands (key STRAND) and turns (key TURN). Residues not specified in one of these classes are in a 'loop' or 'random-coil' structure.

7.36.	Feature Key	UNSURE
	Definition	Uncertainties in the sequence
	Optional qualifiers	note
	Comment	Used to describe region(s) of an amino acid sequence for which the authors are unsure about the sequence presentation.

7.37.	Feature Key	VARIANT
	Definition	Authors report that sequence variants exist
	Optional qualifiers	note

7.38.	Feature Key	VAR_SEQ
	Definition	Description of sequence variants produced by alternative splicing, alternative promoter usage, alternative initiation and ribosomal frameshifting
	Optional qualifiers	note

7.39.	Feature Key	ZN_FING
	Definition	Extent of a zinc finger region
	Mandatory qualifiers	note

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

SECTION 8 : QUALIFICATEURS POUR LES SÉQUENCES D'ACIDES AMINÉS

La présente section donne la liste des qualificateurs pouvant être utilisés pour les séquences d'acides aminés.

Lorsque le format de valeur est du texte libre indiqué comme dépendant de la langue, un des éléments ci-après doit être utilisé :

1) l'élément `INSDQualifier_value`; ou
2) l'élément `NonEnglishQualifier_value`; ou
3) à la fois l'élément `INSDQualifier_value` et l'élément `NonEnglishQualifier_value`.

Lorsque le format de valeur ne constitue pas du texte libre dépendant de la langue, il faut utiliser l'élément `INSDQualifier_value` et il ne faut pas utiliser l'élément `NonEnglishQualifier_value`.

N.B. : Toute valeur de qualificateur indiquée pour un qualificateur avec un format de valeur de texte libre dépendant de la langue peut devoir être traduite aux fins des procédures internationales, nationales ou régionales. Les qualificateurs indiqués dans le tableau ci-après sont considérés comme ayant des valeurs de texte libre dépendant de la langue :

Tableau 6 : liste des valeurs de qualificateurs avec des valeurs de texte libre dépendant de la langue pour les séquences d'acides aminés

Section	Qualificateur de texte libre dépendant de la langue
8.3	note
6.5	organism

8.1.	Qualifier	mol_type
	Definition	In vivo molecule type of sequence
	Mandatory value format	protein
	Example	<INSDQualifier_value>protein</INSDQualifier_value>
	Comment	The "mol_type" qualifier is mandatory on the source feature key.

8.2.	Qualifier	note
	Definition	Any comment or additional information
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>Heme (covalent)</INSDQualifier_value>
	Comment	The "note" qualifier is mandatory for the feature keys: BINDING; CARBOHYD; CROSSLNK; DISULFID; DNA_BIND; DOMAIN; LIPID; METAL; MOD_RES; NP_BIND; SITE and ZN_FING

8.3.	Qualifier	organism
	Definition	Scientific name of the organism that provided the peptide
	Mandatory value format	free text Language-dependent: this value may require translation for International/National/Regional procedures
	Example	<INSDQualifier_value>Homo sapiens</INSDQualifier_value>
	Comment	The "organism" qualifier is mandatory for the source feature key.

SECTION 9 : TABLEAUX DU CODE GÉNÉTIQUE

Le tableau 7 reproduit les tableaux du code génétique à utiliser pour la traduction des séquences de codage. La valeur du qualificateur trans_table est le numéro attribué au tableau du code génétique correspondant. Lorsqu'une caractéristique CDS est décrite par un qualificateur de type traduction, mais non un qualificateur transl_table, on utilise par défaut le 1 – Code normalisé pour la traduction. (Note : les tableaux du code génétique 7, 8, 15 et 17 à 20 n'existant pas, ces numéros n'apparaissent pas dans le tableau 7.)

Tableau 7 : Tableaux du code génétique

1 – Code normalisé

```
  AAs = FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = ---M---------------M---------------M----------------------------
Base1 = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2 = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3 = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

2 – Code mitochondrial des vertébrés

```
  AAs = FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSS**VVVVAAAADDEEGGGG
Starts = --------------------------------MMMM---------------M------------
Base1 = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2 = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3 = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

3 – Code mitochondrial des levures

```
  AAs = FFLLSSSSYY**CCWWTTTTPPPPHHQQRRRRIIMMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = ----------------------------------MM---------------M------------
Base1 = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2 = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3 = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

4 – Code mitochondrial des moisissures, des protozoaires et des cœlentérés et code des mycoplasmes/spiroplasmes

```
  AAs = FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = --MM---------------M------------MMMM---------------M------------
Base1 = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2 = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3 = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

5 – Code mitochondrial des invertébrés

```
  AAs = FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSSSSVVVVAAAADDEEGGGG
Starts = ---M----------------------------MMMM---------------M------------
Base1 = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2 = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3 = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

6 – Code génétique nucléaire des ciliés, des dasycladacées et de l'hexamita

```
  AAs = FFLLSSSSYYQQCC*WLLLLPPPPHHQQRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = -----------------------------------M----------------------------
Base1 = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2 = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3 = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

9 – Code mitochondrial des échinodermes et des platodes

```
  AAs =  F F L L S S S S Y Y * * C C W W L L L L P P P P H H Q Q R R R R I I I M T T T T N N N K S S S S V V V V A A A A D D E E G G G G
Starts = ----------------------------------M---------------M------------
Base1 =  t t t t t t t t t t t t t t t t c c c c c c c c c c c c c c c c a a a a a a a a a a a a a a a a g g g g g g g g g g g g g g g g
Base2 =  t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g
Base3 =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

10 – Code génétique nucléaire des euplotides

```
  AAs =  F F L L S S S S Y Y * * C C C W L L L L P P P P H H Q Q R R R R I I I M T T T T N N K K S S R R V V V V A A A A D D E E G G G G
Starts = ----------------------------------M----------------------------
Base1 = t t t t t t t t t t t t t t t t c c c c c c c c c c c c c c c c a a a a a a a a a a a a a a a a aaggggggggggggggg
Base2 =  t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g
Base3 =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

11 – Code des plastides, des bactéries, des archées et des végétaux

```
  AAs =  F F L L S S SSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = ---M---------------M------------MMMM---------------M------------
Base1 =  t t t t t t t t t t t t t t t t c c c c c c c c c c c c c c c c a a a a a a a a a a a a a a a a g g g g g g g g g g g g g g g g
Base2 =  t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c cccaaaaggggttttccccaaaagggg
Base3 =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

12 – Code génétique nucléaire alternatif des levures

```
  AAs =  F F L L S S S S Y Y * * C C * W L L L S P P P P H H Q Q R R R R I I I M T T T T N N K K S S R R V V V V A A A A D D E E G G G G
Starts = ------------------M---------------M----------------------------
Base1 =  t t t t t t t t t t t t t t t t c c c c c c c c c c c c c c c c a a a a a a a a a a a a a a a a g g g g g g g g g g g g g g g g
Base2 =  t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g
Base3 =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

13 – Code mitochondrial des ascidiens

```
  AAs =  F F L L S S S S Y Y * * C C W W L L L L P P P P H H Q Q R R R R I I M M T T T T N N K K S S G G V V V V A A A A D D E E G G G G
Starts = ---M------------------------------MM---------------M------------
Base1 = t t t t t t t t t t t t t t t t c c c c c c c c c c c c c c c a aaaaaaaaaaaaaaagggggggggggggggg
Base2 =  t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g
Base3 =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

14 – Code mitochondrial alternatif des platodes

```
  AAs =  F F L L S S S S Y Y Y * C CWWLLLLPPPPHHQQRRRRIIIMTTTTNNNKSSSSVVVVAAAADDEEGGGG
Starts = ----------------------------------M----------------------------
Base1 =  t t t t t t t t t t t t t t t t c c c c c c c c c c c c c c c c a a a a a a a a a a a a a a a a g g g g g g g g g g g g g g g g
Base2 =  t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a ggggttttccccaaaagggg
Base3 =  t c a g t c a g t c a g t c a g t c a g t c a g t c a g t c a g t c a g t c a g t c a g t c a g t c a g t c a g t c a g t c a g
Starts = ----------------------------------M----------------------------
Base1 =  t t t t t t t t t t t t t t t t c c c c c c c c c c c c c c c c a a a a a a a a a a a a a a a a g g g g g g g g g g g g g g g g
Base2 =  t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g
Base3 =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

16 – Code mitochondrial des chlorophycées

```
  AAs =  F F L L S S S S Y Y * L C C * W L L L L P P P P H H Q Q R R R R I I I M T T T T N N K K S S R R V V V V A A A A D D E E G GGG
Starts = ----------------------------------M----------------------------
Base1 =  t t t t t t t t t t t t t t t t c c c c c c c c c c c c c c c c a a a a a a a a a a a a a a a a g g g g g g g g g g g g g g g g
Base2 =  t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g
Base3 =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

21 – Code mitochondrial des trématodes

```
  AAs =  F F L L S S S S Y Y * * C C W W L L L L P P P P H H Q Q R R R R I I I M M T T T T N N N K K S S S S V V V V A A A A D D E E G G G G
Starts = ----------------------------------M---------------M------------
Base1 =  t t t t t t ttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2 =  tttt cccc aaaa gggg tttt cccc aaaa gggg tttt cccc aaaa gggg tttt cccc aaaa gggg
Base3 =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

22 – Code mitochondrial de *Scenedesmus obliquus*

```
  AAs =  F F L L S S * S Y Y * L C C * W L L L L P P P P H H Q Q R R R R I I I M T T T T N N K K S S R R V V V V A A A A D D E E G G G G
Starts = ----------------------------------M----------------------------
Base1 =  tttttttttttttttttcccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2 =  t t t t c c c c a a a a g g g g t tttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
Base3 =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

23 – Code mitochondrial de *Thraustochytrium*

```
  AAs =  F F * L S S S S Y Y * * C C * W L L L L P P P P H H Q Q R R R R I I I M T T T T N N K K S S R R V V V V A A A A D D E E G G G G
Starts = ----------------------------------M--M---------------M------------
Base1 =  tttttttttttttttttcccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2 =  tttt cccc aaaa gggg tttt cccc aaaa gggg tttt cccc aaaa gggg tttt cccc aaaa gggg
Base3 =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

24 – Code mitochondrial des ptérobranches

```
  AAs =  F F L L S S S S Y Y * * C C W W L L L L P P P P H H Q Q R R R R I I I M T T T T N N K K S S S K V V V V A A A A D D E E G G G G
Starts = ---M---------------M---------------M---------------M------------
Base1 =  t t t t t t t t t t t t t t t t t c c ccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2 =  tttt cccc aaaa gggg tttt cccc aaaa gggg tttt cccc aaaa gggg tttt cccc aaaa gggg
Base3 =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

25 – Code des bactéries de la Candidate Division SR1 et des bactéries *Gracilis*

```
  AAs  =  FFLLSSSSYY**CCGWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = ---M---------------M---------------M----------------------------
Base1  =  ttttttttttttttttcccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2  =  ttt tccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
Base3  = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

26 – Code génétique nucléaire du champignon *Pachysolen tannophilus*

```
  AAs  = FFLLSSSSYY**CC*WLLLAPPPPHHQQRRRRIIIMTTTTNN KKSSRRVVVVAAAADDEEGGGG
Starts = ---M----------------------------------M---------------M------------
Base1  =  ttttttttttttttttcccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
Base3  = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

27 – Code génétique nucléaire des ciliés de la classe *Karyorelictea*

```
  AAs = F F L L S S S S Y Y Q Q C C W W L L L L P P P P H H Q Q R R R R I I I M T T T T N N K K S S R R V V V V A A A A D D E E G G G G
Starts = ---------------*--------------------M----------------------------
Base1 =  tttttttttttttttttcccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2 =  tttt cccc aaaa gggg tttt cccc aaaa gggg tttt cccc aaaa gggg tttt cccc aaaa gggg
Base3 =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

28 – Code génétique nucléaire des condylostomes

```
  AAs = F F L L S S S S Y Y Q Q C C W W L L L L P P P P H H Q Q R R R R I I I M T T T T N N K K S S R R V V V V A A A A D D E E G G G G
Starts = ----------**--*--------------------M----------------------------
Base1 =  tttttttttttttttttcccccccccccccccaaaaaaaaaaaaaaaag ggggggggggggg
Base2 =  tttt cccc aaaa gggg tttt cccc aaaa gggg tttt cccc aaaa gggg tttt cccc aaaa gggg
Base3 =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

29 – Code génétique nucléaire des ciliés du genre *Mesodinium*

```
  AAs =   F F L L S S S S Y Y Y Y C C * W LLLLPPPPHHQQRRRRIIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = ----------------------------------M----------------------------
Base1 =  t t t t t t t t t t t t t t t t c c c c c c c c c c c c c c c c a a a a a a a a a a a a a a a a g g g g g g g g g g g g g g g g
Base2 =   t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g
Base3 =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

30 – Code génétique nucléaire des péritriches

```
  AAs = F F L L S S S S Y Y E E C C * WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = ----------------------------------M----------------------------
Base1 =  t t t t t t t t t t t t t t t t c c c c c c c c c c c c c c c c a a a a a a a a a a a a a a a a g g g g g g g g g g g g g g g g
Base2 =  t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g
Base3 =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

31 – Code génétique nucléaire du genre *Blastocrithidia*

```
  AAs = F F L L S S S S Y Y E E C C W W L L L L P P P P H H Q Q R R R R I I I M T T T T N N K K S S R R V V V V A A A A D D E E G G G G
Starts = ----------**----------M----------
Base1 = t t t t t t t t t t t t t t t t c c c c c c c c c c c c c c c c a a a a a a a a a a a a a a a a g g g g g g g g g g g g g g g g
Base2 =  t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g t t t t c c c c a a a a g g g g
Base3 =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

33 - Code mitochondrial UAA-Tyr des vers marins *Cephalodiscidae*

```
  AAs   = FFLLSSSSYYY*CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSSKVVVVAAAADDEEGGGG
Start s = -M--- ----------M---------- ----------M---------- ----------M---------- ----------
Base 1 =  ttt ttttt ttt tttt tcccc ccc cccc ccccc aaa aaaa aaaaa aaa aggg ggggg ggg gggg g
Base 2 = ttt tcccc aaa aggg gtttt ccc caaa aggg g ttt tccc caaaa ggg gttt tcccc aaa aggg g
Base 3 =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

[L'annexe II suit]

ANNEXE II

DÉFINITION DE TYPE DE DOCUMENT (DTD) POUR LE LISTAGE DES SÉQUENCES

Version 1. 7

Révision approuvée au Comité des normes de l'OMPI (CWS) à sa onzième session le 8 décembre 2023

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

This entity may be identified by the PUBLIC identifier:
**********************************************************************************************
PUBLIC "-//WIPO//DTD SEQUENCE LISTING 1. 3//EN" "ST26SequenceListing_V1_3.dtd"
**********************************************************************************************
* PUBLIC DTD URL

* https://www.wipo.int/standards/dtd/ST26SequenceListing_V1_3.dtd
**********************************************************************************

*Revision of Annex II to WIPO Standard ST.26 was approved by the Committee on WIPO Standards
*(CWS) at its tenth session.

**********************************************************************************
* CONTACTS
**********************************************************************************
*
*xml.standards@wipo.int
*
**********************************************************************************
* NOTES
**********************************************************************************
*
* The sequence data part is a subset of the complete INSDC DTD V.1.5 that only
*covers the requirements of WIPO Standard ST.26.
*
**********************************************************************************
* REVISION HISTORY
**********************************************************************************
2022-11-25: Comment related to filename approved at CWS/10 (no update to version number)
2021-11-05: Revised Version 1.3 approved at CWS/9 (small edits to the comments)
2020-05-20: Version 1.3 DRAFT.
Changes:
- Optional originalFreeTextLanguageCode attribute added to <ST26SequenceListing> to allow
applicants to indicate the language of the free text in the original sequence listing.
- Optional nonEnglishFreeTextLanguageCode attribute added to <ST26SequenceListing> to allow
applicants to indicate the language of the free text provided in the element
<NonEnglishQualifier_value>.
- Optional id attribute added to INSDQualifier to facilitate comparison of language-
dependent qualifier values between sequence listings.
- Optional element <NonEnglishQualifier_value> added to element <INSDQualifier> to allow
applicants to type language-dependent qualifiers in a non-English Language with the
characters set forth in paragraph 40(a) of the ST.26 main body document.
2018-10-19: Version 1.2 approved at CWS/6.
Changes:
<INSDQualifier*> changed to <INSDQualifier+> for alignment with business needs and advice
from NCBI (an INSDFeature_quals element (if present) should have one or more INSDQualifier
elements).
2017-06-02: Version 1.1 approved at the CWS/5
Changes:
Comments added to <INSDSeq_length>, <INSDSeq_division> and <INSDSeq_sequence> to clarify
the reason of the differences between the INSDC DTD v.1.5 and ST26 Sequence Listing DTD
V1_1.
```

```
2016-03-24: Version 1.0 adopted at the CWS/4Bis
2014-03-11: Final draft for adoption.
********************************************************************************

ST26SequenceListing


********************************************************************************
* ROOT ELEMENT
********************************************************************************
-->
<!ELEMENT ST26SequenceListing ((ApplicantFileReference | (ApplicationIdentification,
ApplicantFileReference?)), EarliestPriorityApplicationIdentification?, (ApplicantName,
ApplicantNameLatin?)?, (InventorName, InventorNameLatin?)?, InventionTitle+,
SequenceTotalQuantity, SequenceData+)>
<!--The elements ApplicantName and InventorName are optional in this DTD to facilitate
the conversion between various encoding schemes-->
<!--originalFreeTextLanguageCode:
The language code (see reference in paragraph 9 to ISO 639-1:2002) for the single original
language in which the language-dependent free text qualifiers (NonEnglishQualifier_value)
were prepared.
-->
<!--nonEnglishFreeTextLanguageCode:
The language code (see reference in paragraph 9 to ISO 639-1:2002) for the language in
which the language-dependent free text qualifiers (NonEnglishQualifier_value) currently
correspond.
-->
<!--fileName:
By default the file name will be set to the value provided for the project name in WIPO
Sequence. If the value is identical to the actual ST.26 XML filename, it should be noted
that Offices may enforce their requirements for the filename used which may restrict which
characters are allowable for submitted electronic files. It is also acceptable for the
value of the filename attribute and the actual file name to be different. Please refer to
the WIPO Sequence and ST.26 Knowledge Base for further details on Offices' naming
conventions for electronic files
--->
<!ATTLIST ST26SequenceListing
               dtdVersion CDATA #REQUIRED
               fileName CDATA #IMPLIED
               softwareName CDATA #IMPLIED
               softwareVersion CDATA #IMPLIED
               productionDate CDATA #IMPLIED
               originalFreeTextLanguageCode CDATA #IMPLIED
               nonEnglishFreeTextLanguageCode CDATA #IMPLIED
>
<!--ApplicantFileReference
Applicant's or agent's file reference, mandatory if application identification not
provided.
-->
<!ELEMENT ApplicantFileReference (#PCDATA)>

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

<!--EarliestPriorityApplicationIdentification
Identification of the earliest priority application, which contains IPOfficeCode,
ApplicationNumberText and FilingDate elements.
-->
<!ELEMENT EarliestPriorityApplicationIdentification (IPOfficeCode, ApplicationNumberText,
             FilingDate?)>

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

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

<!--ApplicantNameLatin
Where ApplicantName is typed in characters other than those as set forth in paragraph 40
(b), a translation or transliteration of the name of the first mentioned applicant must
also be typed in characters as set forth in paragraph 40 (b) of the ST.26 main body
document.
-->
<!ELEMENT ApplicantNameLatin (#PCDATA)>

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

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

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

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

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

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

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

<!--FilingDate
The date of filing of the patent application for which the sequence listing is submitted in
ST.2 format "CCYY-MM-DD", using a 4-digit calendar year, a 2-digit calendar month and a 2-
digit day within the calendar month, e.g., 2015-01-31. For details, please see paragraphs 7
```

```
(a) and 11 of WIPO Standard ST.2.
-->
<!ELEMENT FilingDate (#PCDATA)>

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

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

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

This part is based on:

The International Nucleotide Sequence Database (INSD) collaboration.

INSDSeq provides the elements of a sequence as presented in the GenBank/EMBL/DDBJ-style
flatfile formats. Not all elements are used here.
-->
<!--INSDSeq
Sequence data. Changed INSD V1.5 DTD elements, INSDSeq_division and INSDSeq_sequence from
optional to mandatory per business requirements.
-->
<!ELEMENT INSDSeq (INSDSeq_length, INSDSeq_moltype, INSDSeq_division,
INSDSeq_other-seqids?, INSDSeq_feature-table?, INSDSeq_sequence)>
<!--INSDSeq_length
The length of the sequence. INSDSeq_length allows only integer.
-->
<!ELEMENT INSDSeq_length (#PCDATA)>
<!--INSDSeq_moltype
Admissible values: DNA, RNA, AA
-->
<!ELEMENT INSDSeq_moltype (#PCDATA)>
<!--INSDSeq_division
Indication that a sequence is related to a patent application. Must be populated with the
value PAT.
-->
<!ELEMENT INSDSeq_division (#PCDATA)>
<!--INSDSeq_other-seqids
In the context of data exchange with database providers, the IPOs should populate for each
sequence the element INSDSeq_other-seqids with one INSDSeqid containing a reference to the
corresponding published patent and the sequence identification.
-->
<!ELEMENT INSDSeq_other-seqids (INSDSeqid?)>
<!--INSDSeq_feature-table
Information on the location and roles of various regions within a particular sequence.
Whenever the element INSDSeq_feature-table is used, it must contain at least one feature.
-->
<!ELEMENT INSDSeq_feature-table (INSDFeature+)>
<!--INSDSeq_sequence
The residues of the sequence. The sequence must not contain numbers, punctuation or
whitespace characters.
-->
<!ELEMENT INSDSeq_sequence (#PCDATA)>
<!--INSDSeqid
Intended for the use of IPOs in data exchange only.
Format:
pat|{office code}|{publication number}|{document kind code}|{Sequence identification
number}

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

Example:
pat|WO|2013999999|A1|123456
```

```
This represents the 123456th sequence from WO patent publication No. 2013999999 (A1)
-->
<!ELEMENT INSDSeqid (#PCDATA)>
<!--INSDFeature
Description of one feature.
-->
<!ELEMENT INSDFeature (INSDFeature_key, INSDFeature_location, INSDFeature_quals?)>
<!--INSDFeature_key
A word or abbreviation indicating a feature.

-->
<!ELEMENT INSDFeature_key (#PCDATA)>
<!--INSDFeature_location
Region of the presented sequence which corresponds to the feature.
-->
<!ELEMENT INSDFeature_location (#PCDATA)>
<!--INSDFeature_quals
List of qualifiers containing auxiliary information about a feature.
-->
<!ELEMENT INSDFeature_quals (INSDQualifier+)>
<!--INSDQualifier
Additional information about a feature.
For coding sequences and variants see the ST.26 main body document.
-->
<!--id
Unique identifier for the INSDQualifier to facilitate comparison of versions of a sequence
listing specifically having language-dependent qualifier values in different languages.
-->
<!ELEMENT INSDQualifier (INSDQualifier_name, INSDQualifier_value?,
NonEnglishQualifier_value?)>
<!ATTLIST INSDQualifier
       id ID #IMPLIED
>
<!--INSDQualifier_name
Name of the qualifier.
-->
<!ELEMENT INSDQualifier_name (#PCDATA)>
<!--INSDQualifier_value
Value of the qualifier. Where the qualifier is language-dependent its value must be in the
English language and typed with the characters set forth in paragraph 40 (b).
-->
<!ELEMENT INSDQualifier_value (#PCDATA)>
<!--NonEnglishQualifier_value
Value of a language-dependent qualifier in a language that is not English and typed with
the characters set forth in paragraph 40 (a). The language is indicated with the attribute
nonEnglishFreeTextLanguageCode.
-->
<!ELEMENT NonEnglishQualifier_value (#PCDATA)>
```

[L'annexe III suit]

ANNEXE III

EXEMPLE DE LISTAGE DES SÉQUENCES (fichier XML)

Version 1.7

Révision approuvée au Comité des normes de l'OMPI (CWS) à sa onzième session le 8 décembre 2023

L'annexe III est disponible à l'adresse suivante :
https://www.wipo.int/standards/en/xml_material/st26/st26-annex-iii-sequence-listing-specimen.xml

[L'annexe IV suit]

ANNEXE IV

SOUS-ENSEMBLE DE CARACTÈRES PROVENANT DU TABLEAU DE CODES DES CARACTÈRES LATINS DE BASE DE LA NORME UNICODE DEVANT ÊTRE EMPLOYÉS DANS L'INSTANCE XML D'UN LISTAGE DE SÉQUENCES

Version 1.7

Révision approuvée au Comité des normes de l'OMPI (CWS) à sa onzième session le 8 décembre 2023

La perluète (0026) n'est autorisée qu'en tant qu'élément d'une entité prédéfinie. Le guillemet (0022), l'apostrophe (0027), le signe inférieur à (003C) et le signe supérieur à (003E) doivent être représentés par leurs entités prédéfinies. En outre, la perluète (0026) doit être présentée par son entité prédéfinie lorsqu'elle est utilisée comme perluète dans une valeur d'un attribut ou contenu d'un élément.

Unicode code point	Character	Name
0020		SPACE
0021	!	EXCLAMATION MARK
0022	"	QUOTATION MARK
0023	#	NUMBER SIGN
0024	$	DOLLAR SIGN
0025	%	PERCENT SIGN
0026	&	AMPERSAND
0027	'	APOSTROPHE
0028	(	LEFT PARENTHESIS
0029	)	RIGHT PARENTHESIS
002A	*	ASTERISK
002B	+	PLUS SIGN
002C	,	COMMA
002D	-	HYPHEN-MINUS
002E	.	FULL STOP
002F	/	SOLIDUS
0030	0	DIGIT ZERO
0031	1	DIGIT ONE
0032	2	DIGIT TWO
0033	3	DIGIT THREE
0034	4	DIGIT FOUR
0035	5	DIGIT FIVE
0036	6	DIGIT SIX
0037	7	DIGIT SEVEN
0038	8	DIGIT EIGHT
0039	9	DIGIT NINE
003A	:	COLON
003B	;	SEMICOLON
003C	<	LESS-THAN-SIGN
003D	=	EQUALS SIGN
003E	>	GREATER-THAN-SIGN
003F	?	QUESTION MARK
0040	@	COMMERCIAL AT
0041	A	LATIN CAPITAL LETTER A
0042	B	LATIN CAPITAL LETTER B
0043	C	LATIN CAPITAL LETTER C
0044	D	LATIN CAPITAL LETTER D
0045	E	LATIN CAPITAL LETTER E
0046	F	LATIN CAPITAL LETTER F
0047	G	LATIN CAPITAL LETTER G
0048	H	LATIN CAPITAL LETTER H
0049	I	LATIN CAPITAL LETTER I

Unicode code point	Character	Name
004B	K	LATIN CAPITAL LETTER K
004C	L	LATIN CAPITAL LETTER L
004D	M	LATIN CAPITAL LETTER M
004E	N	LATIN CAPITAL LETTER N
004F	O	LATIN CAPITAL LETTER O
0050	P	LATIN CAPITAL LETTER P
0051	Q	LATIN CAPITAL LETTER Q
0052	R	LATIN CAPITAL LETTER R
0053	S	LATIN CAPITAL LETTER S
0054	T	LATIN CAPITAL LETTER T
0055	U	LATIN CAPITAL LETTER U
0056	V	LATIN CAPITAL LETTER V
0057	W	LATIN CAPITAL LETTER W
0058	X	LATIN CAPITAL LETTER X
0059	Y	LATIN CAPITAL LETTER Y
005A	Z	LATIN CAPITAL LETTER Z
005B	[	LEFT SQUARE BRACKET
005C	\	REVERSE SOLIDUS
005D	]	RIGHT SQUARE BRACKET
005E	^	CIRCUMFLEX ACCENT
005F	_	LOW LINE
0060	`	GRAVE ACCENT
0061	a	LATIN SMALL LETTER A
0062	b	LATIN SMALL LETTER B
0063	c	LATIN SMALL LETTER C
0064	d	LATIN SMALL LETTER D
0065	e	LATIN SMALL LETTER E
0066	f	LATIN SMALL LETTER F
0067	g	LATIN SMALL LETTER G
0068	h	LATIN SMALL LETTER H
0069	i	LATIN SMALL LETTER I
006A	j	LATIN SMALL LETTER J
006B	k	LATIN SMALL LETTER K
006C	l	LATIN SMALL LETTER L
006D	m	LATIN SMALL LETTER M
006E	n	LATIN SMALL LETTER N
006F	o	LATIN SMALL LETTER O
0070	p	LATIN SMALL LETTER P
0071	q	LATIN SMALL LETTER Q
0072	r	LATIN SMALL LETTER R
0073	s	LATIN SMALL LETTER S
0074	t	LATIN SMALL LETTER T
0075	u	LATIN SMALL LETTER U
0076	v	LATIN SMALL LETTER V
0077	w	LATIN SMALL LETTER W
0078	x	LATIN SMALL LETTER X
0079	y	LATIN SMALL LETTER Y
007A	z	LATIN SMALL LETTER Z
007B	{	LEFT CURLY BRACKET
007C	|	VERTICAL LINE
007D	}	RIGHT CURLY BRACKET
007E	~	TILDE

[L'annexe V suit]

ANNEXE V

PRESCRIPTIONS SUPPLÉMENTAIRES EN MATIÈRE D'ÉCHANGE DE DONNÉES (UNIQUEMENT POUR LES OFFICES DE PROPRIÉTÉ INTELLECTUELLE)

Version 1.7

Révision approuvée au Comité des normes de l'OMPI (CWS) à sa onzième session le 8 décembre 2023

Dans le contexte de l'échange de données avec les fournisseurs de bases de données (membres de la Collaboration internationale sur les bases de données de séquences de nucléotides (INSDC)), les offices de propriété intellectuelle doivent, pour chaque séquence, insérer dans l'élément `INSDSeq_other-seqids` un élément `INSDSeqid` contenant une référence au brevet publié correspondant et le numéro d'identification de la séquence selon le format ci-après :

brevet |{code de l'office }|{numéro de publication }|{code du type de document}|{numéro d'identification de la séquence}

où le code de l'office est le code de l'office de la propriété intellectuelle qui a publié le document de brevet conformément à la norme ST.3; le code du type de document est le code d'identification des différents types de documents de brevet conformément à la norme ST.16; le numéro de publication est le numéro de publication de la demande ou du brevet; et le numéro d'identification de la séquence est le numéro de la séquence indiqué dans cette demande ou ce brevet.

Exemple :

brevet|WO|2013999999|A1|123456

qui se traduirait dans l'instance XML valide qui suit :

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

où "123456" est la 123456^{e} séquence provenant du numéro de publication WO 2013999999 (A1).

[L'annexe VI suit]

ANNEXE VI

GUIDE D'APPLICATION ASSORTI D'EXEMPLES ILLUSTRATIFS

Version 1.7

Révision approuvée au Comité des normes de l'OMPI (CWS) à sa onzième session le 8 décembre 2023

TABLE DES MATIERES

INTRODUCTION

La présente norme a notamment pour but 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". Le présent document d'orientation vise à faire en sorte que tous les déposants et offices de la propriété intellectuelle comprennent et acceptent les exigences en matière d'intégration et de représentation des divulgations de séquences, de sorte que ce but puisse être atteint.

Le présent document comprend la présente introduction, un index des exemples, des exemples de divulgation de séquences et un appendice contenant un listage des séquences en XML établi à partir des séquences tirées des exemples. La présente introduction explique certains concepts et termes utilisés dans le reste du document. Les exemples illustrent les exigences des différents paragraphes de la norme et il est renvoyé pour chaque exemple au numéro de paragraphe qui correspond le mieux à l'objet traité. Par ailleurs, certains exemples illustrent d'autres paragraphes, et des renvois appropriés sont indiqués à la fin de chaque exemple. L'index présente les numéros de page des exemples, avec indications des éventuels renvois. Chaque séquence d'un exemple qui doit ou peut être intégrée dans un listage des séquences s'est vu attribuer un numéro d'identification de séquence (SEQ ID NO) et apparaît au format XML dans l'Appendice au présent document.

Pour chaque exemple, toute information donnée à titre d'explication d'une séquence doit être considérée comme l'intégralité de la divulgation concernant cette séquence. Les réponses données ne tiennent compte que des informations présentées de manière explicite dans l'exemple.

Les orientations fournies dans le présent document portent sur l'établissement d'un listage des séquences à fournir à la date du dépôt d'une demande de brevet. Quant à l'établissement d'un listage des séquences à fournir après la date du dépôt d'une demande de brevet, il convient de se demander si les informations fournies pourraient être considérées par un office de la propriété intellectuelle comme ajoutant des éléments à la divulgation originale. Il s'ensuit que les orientations fournies dans le présent document pourront ne pas être applicables à un listage des séquences présenté après la date du dépôt d'une demande de brevet.

Établissement d'un listage des séquences

Pour établir un listage des séquences, il importe de se poser les questions suivantes :

1. Le paragraphe 7 de la norme ST.26 prescrit-il l'intégration d'une séquence divulguée donnée?

2. Si l'intégration d'une séquence divulguée donnée n'est pas prescrite, l'intégration de cette séquence est-elle autorisée par la norme ST.26?

3. Si l'intégration d'une séquence divulguée donnée est prescrite ou autorisée par la norme ST.26, comment cette séquence doit-elle être représentée dans le listage des séquences?

S'agissant de la première question, le paragraphe 7 (avec certaines restrictions) de la norme ST.26 prescrit l'intégration d'une séquence qui est divulguée dans une demande de brevet par l'énumération de ses résidus, lorsque la séquence contient au moins 10 nucléotides définis de manière spécifique ou au moins quatre acides aminés définis de manière

spécifique.

S'agissant de la deuxième question, le paragraphe 8 de la norme ST.26 interdit l'intégration d'une séquence comportant moins de 10 nucléotides définis de manière spécifique ou moins de quatre acides aminés définis de manière spécifique.

Pour répondre à ces deux questions, il importe de bien comprendre l'"énumération de ses résidus" et "définis de manière spécifique".

S'agissant de la troisième question, le présent document contient les divulgations de séquences qui illustrent différents scénarios, ainsi qu'un examen exhaustif du moyen de représentation préféré de chaque séquence ou, lorsqu'une séquence présente de multiples variantes, de la "séquence la plus englobante", conformément à la présente norme. Vu l'impossibilité de prendre en compte tous les scénarios de séquences inhabituelles possibles, le présent document d'orientation s'efforce de présenter le raisonnement qui sous-tend l'approche de chaque exemple et la manière dont les dispositions de la norme ST.26 sont appliquées, de sorte que le même raisonnement peut être appliqué aux autres scénarios de séquences non illustrés.

Énumération de ses résidus

Le paragraphe 3.c) de la norme ST.26 définit l'expression "énumération de ses résidus" comme désignant la divulgation d'une séquence dans une demande de brevet sous forme de listage, dans un ordre donné, de chacun des résidus de la séquence, étant entendu que i) le résidu est représenté par un nom, une abréviation, un symbole ou une structure; ou ii) les résidus multiples sont représentés par une formule topologique. Une séquence devrait être divulguée dans une demande de brevet par "l'énumération de ses résidus" à l'aide de symboles conventionnels, qui sont les symboles des nucléotides indiqués dans le tableau 1 de la section 1 de l'annexe 1 à la norme ST.26 (c'est-à-dire les symboles en lettres minuscules ou leurs équivalents en lettres majuscules[1]) et les symboles des acides aminés indiqués dans le tableau 3 de la section 3 de l'annexe 1 à la norme ST.26 (c'est-à-dire les symboles en lettres majuscules ou leurs équivalents en lettres minuscules[1]). Les symboles de ces nucléotides et de ces acides aminés sont ci-après dénommés symboles conventionnels. Les représentations de ces nucléotides et acides aminés autres que celles indiquées dans ces tableaux sont dénommées symboles "non conventionnels".

Lorsqu'une représentation d'un résidu est divulgué comme équivalant à un symbole ou une abréviation conventionnel (par exemple, "Z1" désigne "A") ou à une séquence particulière de symboles conventionnels (par exemple, "Z1" désigne "agga"), la séquence est interprétée comme si elle était divulguée à l'aide du ou des symboles ou abréviations conventionnels équivalents, afin de déterminer si le paragraphe 7 de la norme ST.26 prescrit l'intégration dans le listage des séquences ou si le paragraphe 8 interdit cette intégration. Lorsqu'un symbole non conventionnel de nucléotide est utilisé comme symbole ambigu (par exemple, X1 = inosine ou pseudouridine), mais sans être l'équivalent de l'un des symboles ambigus conventionnels indiqués dans le tableau 1 de la section 1 (c'est-à-dire "m", "r", "w", "s", "y", "k", "v", "h", "d", "b", ou "n"), le résidu est interprété en tant que résidu "n" afin de déterminer si le paragraphe 7 de la norme ST.26 prescrit l'intégration dans le listage des séquences ou si le paragraphe 8 de la norme ST.26 interdit cette intégration. De même, lorsqu'un symbole non conventionnel d'acide aminé est utilisé comme symbole ambigu (par exemple "Z1" désigne "A", "G", "S" ou "T"), mais sans être l'équivalent de l'un des symboles ambigus conventionnels indiqués dans le tableau 3 de la section 3 (c'est-à-dire B, Z, J ou X), le résidu est interprété en tant que résidu "X" afin de déterminer si le paragraphe 7 de la norme ST.26 prescrit l'intégration de la séquence dans le listage des séquences ou si le paragraphe 8 de la norme ST.26 interdit cette intégration.

Il faudrait s'efforcer de divulguer les séquences à l'aide de symboles conventionnels; toutefois, lorsque les séquences sont divulguées autrement, il peut être nécessaire de consulter l'explication de la séquence donnée dans la divulgation afin de déterminer la signification de la représentation non conventionnelle.

Lorsqu'un symbole conventionnel est utilisé, on n'en doit pas moins consulter l'explication de la séquence donnée dans la divulgation afin de confirmer que le symbole est utilisé d'une manière conventionnelle. Si le symbole est utilisé d'une manière non conventionnelle, cette explication est nécessaire pour déterminer si le paragraphe 7 de la norme ST.26 prescrit l'intégration de la séquence dans le listage des séquences ou si le paragraphe 8 interdit cette intégration.

Définis de manière spécifique

Le paragraphe 3.k) de la norme ST.26 définit l'expression "définis de manière spécifique" comme désignant 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, étant entendu que "n" et "X" sont utilisés d'une manière conventionnelle telle que décrite dans le tableau 1 de la section 1 (c'est-à-dire "a ou c ou g ou t/u; 'unknown' ou 'other'") et dans le tableau 3 de la section 3 (c'est-à-dire A ou R ou N ou D ou C ou Q ou E ou G ou H ou I ou L ou K ou M ou F ou P ou O ou S ou U ou T ou W ou Y ou V, 'unknown' ou 'other'"), respectivement. Pour déterminer si un nucléotide ou un acide aminé est "défini de manière spécifique", il sera tenu compte des indications susmentionnées concernant les symboles conventionnels ou les symboles ou abréviations non conventionnels et leur utilisation d'une manière conventionnelle ou non conventionnelle.

Séquence la plus englobante

Lorsqu'une séquence qui répond aux exigences du paragraphe 7 n'est divulguée par l'énumération de ses résidus qu'une seule fois dans une demande, mais est décrite d'une manière différente dans de multiples modes de réalisation – par exemple, dans un mode de réalisation, "X" pourrait être, à un ou plusieurs emplacements, tout acide aminé, mais, dans des

modes de réalisation ultérieurs, "X" pourrait ne représenter qu'un nombre limité d'acides aminés –, la norme ST.26 prescrit

[1] NOTE : Si les divulgations jointes à une demande peuvent représenter les nucléotides ou les acides aminés par des symboles en lettres minuscules ou majuscules, lorsqu'il s'agit d'une séquence intégrée dans un listage des séquences, seules les lettres minuscules sont autorisées pour représenter une séquence de nucléotides (voir le paragraphe 13 de la norme ST.26) et seules les lettres majuscules le sont pour représenter une séquence d'acides aminés (voir le paragraphe 26 de la norme ST.26).

l'intégration dans un listage des séquences uniquement de la séquence qui a été énumérée par ses résidus. Conformément aux paragraphes 15 et 27, lorsqu'une telle séquence contient des symboles ambigus multiples "n" ou "X", "n" ou "X" est considéré comme représentant tout nucléotide ou acide aminé, respectivement, en l'absence d'annotation supplémentaire. En conséquence, la seule séquence à intégrer est la séquence la plus englobante divulguée. La séquence la plus englobante est la séquence unique dont les résidus de variante sont représentés par les symboles ambigus les plus restrictifs qui intègrent les modes de réalisation les plus divulgués. De même, lorsqu'une séquence est divulguée par l'énumération de ses résidus qu'une seule fois, mais que la longueur de la séquence peut varier en raison de la variation du nombre de copies de la répétition, le mode de réalisation le plus long est considéré comme la séquence la plus englobante. Prenons, par exemple, une séquence contenant une région répétée qui peut varier de 2 à 5 copies telles qu'énumérées. Le mode de réalisation comprenant 5 copies de la répétition est la séquence la plus englobante et devrait être intégrée dans le listage des séquences. Toutefois, l'intégration de séquences particulières supplémentaires est fortement conseillée lorsqu'elle est possible, par exemple celles qui représentent des modes de réalisation supplémentaires qui constituent une partie essentielle de l'invention. L'intégration des séquences supplémentaires permet d'effectuer une recherche plus approfondie et porte à la connaissance du public l'objet pour lequel la délivrance d'un brevet est demandée.

Usage du symbole ambigu

Bon usage du symbole ambigu "n" dans un listage des séquences

Le symbole "n"

a. ne doit être employé que pour représenter un seul nucléotide;

b. sera considéré comme l'équivalent de l'un des symboles "a", "c", "g" ou "t/u", sauf s'il est accompagné d'une description supplémentaire;

c. devrait être employé pour représenter l'un quelconque des nucléotides ci-après s'il est accompagné d'une description supplémentaire :

i. un nucléotide modifié, par exemple un nucléotide naturel, synthétique ou n'existant pas à l'état naturel, qui ne peut être représenté à l'aide d'un autre symbole indiqué dans l'annexe I (voir section 1, tableau 1);

ii. un nucléotide "unknown", c'est-à-dire non déterminé, non divulgué ou incertain;

iii. un site abasique; ou

d. peut être employé pour représenter une variante de séquence, c'est-à-dire des alternatives, des suppressions, des adjonctions ou des remplacements, lorsque "n" est le symbole ambigu le plus restrictif.

Bon usage du symbole ambigu "X" dans un listage des séquences

Le symbole "X"

a. ne peut être employé que pour représenter un acide aminé.

b. 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

c. devrait être employé pour représenter l'un quelconque des acides aminés ci-après s'il est accompagné d'une description supplémentaire :

i. un acide aminé modifié, par exemple un acide aminé naturel, synthétique ou n'existant pas à l'état naturel, qui ne peut être représenté à l'aide d'un autre symbole indiqué dans l'annexe I (voir section 3, tableau 3);

ii. un acide aminé "unknown", c'est-à-dire non déterminé, non divulgué ou incertain; ou

d. peut être employé pour représenter une variante de séquence, c'est-à-dire des alternatives, des

suppressions, des adjonctions ou des remplacements, lorsque "X" est le symbole ambigu le plus restrictif.

Annotation des résidus modifiés

La présente norme exige que les résidus "modifiés" soient annotés conformément au paragraphe 17 pour les nucléotides et conformément au paragraphe 30 pour les acides aminés.

Le paragraphe 3.e) de la norme ST.26 définit un "acide aminé modifié" comme tout acide aminé tel que décrit au paragraphe 3.a) différent de L-alanine, L-arginine, L-asparagine, L-aspartate, L-cystéine, L-glutamine, L-glutamate, L-glycine, L-histidine, L-isoleucine, L-leucine, L-lysine, L-méthionine, L-phénylalanine, L-proline, L-pyrrolysine, L-sérine, L-sélénocystéine, L-thréonine, L-tryptophane, L-tyrosine ou L-valine. De même, la norme définit un "nucléotide modifié" comme tout nucléotide tel que décrit au paragraphe 3.g), différent de la désoxyadénosine 5'-monophosphate, de la désoxyguanosine 5'-monophosphate, de la désoxycytidine 5'-monophosphate, de la désoxythymidine 5'-monophosphate, de l'adénosine 5'-monophosphate, de la guanosine 5'-monophosphate, de la cytidine 5'-monophosphate ou de l'uridine 5'-monophosphate (norme ST.26, paragraphe 3.f)).

Compte tenu des définitions ci-dessus, les modifications des bases azotées ou du squelette sucre-phosphate d'un acide nucléique et les modifications des groupes R d'acides aminés ou d'un squelette de peptide engendrent un ou plusieurs "nucléotides modifiés" ou "acides aminés modifiés", respectivement. Par conséquent, de tels nucléotides et acides aminés doivent être annotés. Parmi les exemples de modifications de squelette, l'on peut citer les analogues nucléotidiques comme les acides nucléiques peptidiques (ANP) et les acides nucléiques à glycol (ANG) ainsi que les acides aminés D.

Il convient de noter que la modification d'un acide aminé terminal d'un peptide ou d'un nucléotide terminal d'un acide nucléique n'engendre pas nécessairement un "acide aminé modifié " ou un "nucléotide modifié". Il faut examiner la modification terminale et déterminer si elle modifie la structure chimique du résidu de telle façon que le résidu sort du cadre des exceptions visées au paragraphe 3.e) et 3.f). Par exemple, un peptide dans lequel un résidu C-terminal est relié à une structure (comme une partie d'une séquence ramifiée – voir le peptide n° 2 dans l'exemple 7(b)-3) via une liaison amide conventionnelle n'est pas considéré comme un "résidu modifié" et n'a donc pas besoin d'être annoté. De même, un peptide dans lequel un résidu N-terminal est relié par une liaison amide à une biotine n'est pas considéré comme un "résidu modifié" et n'a donc pas besoin d'être annoté. Dans ces deux cas de figure, la structure du résidu impliqué dans la liaison du C-terminal et du N-terminal n'est pas modifiée par rapport aux acides aminés conventionnels mentionnés au paragraphe 3.e) de la norme.

En revanche, les modifications terminales qui changent la structure chimique du résidu sont considérées comme des "résidus modifiés" et doivent être annotées. Par exemple, la méthylation de l'extrémité C-terminale dans l'exemple 3(c)-1 modifie la structure chimique du résidu terminal, puisque le groupe méthyle remplace l'hydroxyle que l'on trouve normalement dans le groupe alpha carboxyle. Par conséquent, ce résidu doit être annoté en tant que "résidu modifié".

Il convient de noter qu'il appartiendra au déposant d'évaluer chaque modification terminale de résidu dans une séquence énumérée et de déterminer si la structure du résidu terminal est ou non modifiée. Si la structure modifiée du résidu est différente des acides aminés ou des nucléotides indiqués au paragraphe 3.e) et 3.f) de la norme, la modification doit être annotée.

Enfin, il est toujours recommandé aux déposants d'inclure autant d'informations que possible dans leurs listages de séquences pour représenter leurs divulgations de manière aussi précise que possible. Par conséquent, même si une modification n'a pas besoin d'être annotée, elle devrait être de préférence incluse.

Il convient cependant de noter que cette annotation des variantes d'une séquence primaire énumérée doit être conforme aux exigences des paragraphes 93 à 100 de la norme ST.26. Les modifications qui sont divulguées en tant que variantes d'une séquence énumérée peuvent ne pas avoir besoin d'être incluses dans le listage des séquences. Pour la définition des annotations de variantes, voir les paragraphes 93 à 95.

Représentation des résidus modifiés

La norme ST.26 indique que les nucléotides et les acides aminés devraient être représentés dans le listage des séquences comme le résidu non modifié correspondant chaque fois que possible (voir les paragraphes 16 et 29). Il convient de noter que cette recommandation est introduite par un "devrait" : "une démarche fortement conseillée, mais pas obligatoire" (voir paragraphe 4.d)). Il est laissé à la discrétion du déposant de décider si un résidu modifié sera représenté par un résidu non modifié correspondant ou par les variables "n" or "X".

En règle générale, si un résidu est modifié par l'ajout d'une fraction, telle que la méthylation ou l'acétylation, et que la structure du résidu non modifié demeure généralement inchangée, la représentation par un résidu non modifié est alors

recommandée. Par exemple, l'adénosine méthylée devrait être représentée par un "a" dans le listage des séquences. Toutefois, lorsque le résidu modifié est structurellement différent de tout résidu non modifié, il est alors recommandé d'utiliser un "n" ou un "X". Par exemple, la norleucine est un isomère de la leucine, et sa chaîne latérale est une structure linéaire de 4 carbones. La leucine comporte également une chaîne latérale de 4 carbones, mais elle est ramifiée au niveau du second carbone. Par conséquent, la norleucine n'est pas simplement le résultat d'une modification ajoutée à la leucine, mais une structure complètement différente (bien qu'apparentée). Il est donc recommandé de représenter la norleucine par un "X" dans un listage des séquences.

Un nucléotide est "spécialement défini" lorsqu'il est représenté par tout symbole différent de "n" et un acide aminé est "spécialement défini" lorsqu'il est représenté par tout symbole différent de "X" (voir le paragraphe 3.k) de la norme ST.26). La 2'-O-méthyladénosine représentée par un 'a' dans la séquence est donc spécialement définie, alors que la norleucine représentée par un 'X' dans la séquence n'est pas spécialement définie.

Tableau A – Symboles conventionnels des nucléotides et définitions

Symbole	**Définition**
a	Adénine
c	Cytosine
g	Guanine
t	Thymine dans l'ADN Uracile dans l'ARN (t/u)
m	a ou c
r	a ou g
w	a ou t/u
s	c ou g
y	c ou t/u
k	g ou t/u
v	a ou c ou g; et non t/u
h	a ou c ou t/u; et non g
d	a ou g ou t/u; et non c
b	c ou g ou t/u; et non a
n	a or c or g or t/u; "unknown" or "other"

Tableau B – Symboles conventionnels des acides aminés, codes de trois lettres et définitions

Symbole	Code de trois lettres	Définition
A	Ala	Alanine
r	Arg	Arginine
n	Asn	Asparagine
d	Asp	Acide aspartique (aspartate)
c	Cys	Cystéine
Q	Gln	Glutamine
E	Glu	Acide glutamique (glutamate)
g	Gly	Glycine
h	His	Histidine
I	Ile	Isoleucine
L	Leu	Leucine
k	Lys	Lysine
m	Met	Méthionine
F	Phe	Phénylalanine
P	Pro	Proline
O	Pyl	Pyrrolysine
s	Ser	Sérine
U	Sec	Sélénocystéine
t	Thr	Thréonine
w	Trp	Tryptophane
y	Tyr	Tyrosine
V	Val	Valine
b	Asx	Acide aspartique ou asparagine
Z	Glx	Glutamine ou acide glutamique
J	Xle	Leucine ou isoleucine
X	Xaa	A ou R ou N ou D ou C ou Q ou E ou G ou H ou I ou L ou K ou M ou F ou P ou O ou S ou U ou T ou W ou Y ou V; "unknown" ou "other"

INDEX DES EXEMPLES

EXEMPLES

Paragraphe 3.a) – Définition d'"acide aminé"

Exemple 3.a)-1 : acides aminés D

Une demande de brevet décrit la séquence ci-après :

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

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Le paragraphe 3.a) de la norme définit "acide aminé" comme comprenant les "acides aminés D" et les acides aminés contenant des chaînes latérales modifiées ou synthétiques. Conformément à cette définition, le peptide énuméré contient cinq acides aminés qui sont définis de manière spécifique (D-Ala, D-Glu, Lys, Gly et D-Met). La séquence doit donc être intégrée dans un listage des séquences comme l'exige le paragraphe 7.b) de la norme ST.26.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Le paragraphe 29 prescrit que les acides aminés modifiés devraient être représentés dans la séquence comme l'acide aminé L non modifié correspondant. Par ailleurs, tout acide aminé modifié ne pouvant être représenté à l'aide d'un autre symbole indiqué dans l'annexe I, section 3, tableau 3 doit être représenté par le symbole "X".

Dans cet exemple, la séquence contient trois acides aminés D qui peuvent être représentés par un acide aminé L non modifié indiqué dans l'annexe I, section 3, tableau 3, un acide aminé L (Nle) et un acide aminé D (D-Nle) qui doit être représenté par le symbole "X".

Aux termes du paragraphe 25, lorsque les séquences d'acides aminés présentent une configuration circulaire et que l'anneau se compose uniquement de résidus d'acides aminés liés par des peptides, le déposant doit choisir l'acide aminé à la position de résidu numéro 1. La séquence peut donc être représentée comme suit :

AEKXGMX (SEQ ID NO : 1)

ou autrement, avec un autre acide aminé de la séquence en position de résidu numéro 1. Une clé de caractérisation "SITE" et un qualificateur "note" doivent être indiqués pour chaque acide aminé D, le nom complet non abrégé de l'acide aminé D étant indiqué dans la valeur du qualificateur, par exemple l'alanine D et la norleucine D. Par ailleurs, une clé de caractérisation "SITE" et un qualificateur "note" doivent être indiqués, le qualificateur prenant comme valeur l'abréviation de la norleucine L, c'est-à-dire "Nle", conformément à ce qui est indiqué dans l'annexe I, section 4, tableau 4. Enfin, une clé de caractérisation "REGION" et un qualificateur "note" devraient être prévus pour montrer que le peptide est circulaire.

Paragraphes pertinents de la norme ST.26 : 3.a), 7.b), 25, 26, 29, 30 et 31

Paragraphe 3.c) – Définition de "énumération de ses résidus"

Exemple 3.c)-1 : Énumération des acides aminés selon la structure chimique

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

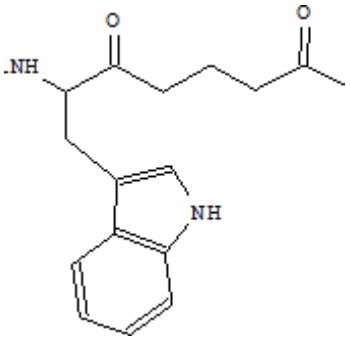
OUI

Le peptide énuméré, illustré sous forme d'une structure, contient au moins quatre acides aminés définis de manière spécifique. La séquence doit donc être intégrée dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence peut être représentée sous la forme suivante :

VAFXGK (SEQ ID NO : 2)



où "X" représente un acide aminé modifié "other", qui nécessite une clé de caractérisation "SITE" ainsi que le qualificateur "note". Le qualificateur "note" indique le nom complet non abrégé du tryptophane modifié à la position 4 du peptide énuméré, par exemple l'"acide 6-amino-7-(1H-indol-3-yl)-5-oxoheptanoïque". La méthylation de l'extrémité C-terminale modifie la structure de la lysine terminale, étant donné que -OH sur l'extrémité terminale est remplacé par -CH3. En raison de cette modification structurelle, la lysine dans la séquence est considérée comme un "acide aminé modifié". Par conséquent, une clé de caractérisation "SITE" et un qualificateur "note" doivent indiquer la méthylation de l'extrémité C-terminale. La valine n'est toutefois pas considérée comme un "acide aminé modifié", étant donné que l'ajout du groupe acétyle à la valine implique une liaison peptidique conventionnelle. L'acétylation ne modifie pas la structure de la valine. Par conséquent, une clé de caractérisation supplémentaire "SITE" et un qualificateur "note" devraient être inclus pour indiquer l'acétylation de l'extrémité N-terminale.

La séquence pourrait également être représentée sous la forme suivante :

VAFW (SEQ ID NO : 3)

Une clé de caractérisation "SITE" et un qualificateur "note" sont requis pour indiquer une modification du tryptophane à la position 4 du peptide énuméré avec la valeur "extrémité C-terminale liée au dipeptide GK par l'intermédiaire d'un pont de glutaraldéhyde". D'autre part, une clé de caractérisation "SITE" à l'emplacement 1 et un qualificateur "note" supplémentaires devraient être inclus pour indiquer l'acétylation de l'extrémité N-terminale.

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.c), 7.b), 29, 30 et 31

Exemple 3.c)-2 : Formule topologique pour une séquence d'acides aminés

$(G_4z)_n$

où G= Glycine, z = tout acide aminé et la variable n peut être tout nombre entier.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

La divulgation indique que "n" peut être "tout nombre entier"; le mode de réalisation le plus englobant de "n" est donc indéterminé. Étant donné que "n" est indéterminé, le peptide de la formule ne peut pas être amplifié jusqu'à une longueur déterminée et, de ce fait, la formule non amplifiée doit être considérée.

Le peptide énuméré de la formule non amplifiée ("n" = 1) contient quatre acides aminés définis de manière spécifique, dont chacun est Gly, et le symbole "z". Par convention, "Z" est le symbole de la "glutamine ou (de l') acide glutamique"; toutefois, l'exemple définit "z" comme "tout acide aminé". Conformément à la norme ST.26, un acide aminé qui n'est pas défini de manière spécifique est représenté par "X". Il découle de cette analyse que le peptide énuméré, c'est-à-dire GGGGX, contient quatre résidus de glycine énumérés et spécifiquement définis. De ce fait, le paragraphe 7.b) de la norme ST.26 prescrit l'intégration de la séquence dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence emploie un symbole non conventionnel "z", dont la définition doit être déterminée à partir de la divulgation (voir l'Introduction au présent document). Étant donné que "z" est défini comme tout acide aminé, le symbole conventionnel servant à représenter cet acide aminé est "X." De ce fait, la séquence doit être représentée comme une séquence unique :

GGGGX (SEQ ID NO : 4)

et devrait être annotée à l'aide de la clé de caractérisation REGION, de l'emplacement de caractéristique ">5" (correspond à >5) et d'un qualificateur note qui prend comme valeur "L'ensemble de la séquence d'acides aminés 1-5 peut être répété une ou plusieurs fois".

Conformément au paragraphe 27, ""sera considéré comme l'équivalent de 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 dans le tableau de caractéristiques. Comme dans cet exemple, "X" représente "tout acide aminé", il doit être annoté au moyen de la clé de caractérisation "VARIANT" et d'un qualificateur note prenant la valeur "X peut être tout acide aminé".

Chaque fois que possible, chaque "X" devrait être annoté individuellement. Cependant, une région de résidus "X" contigus ou un grand nombre de résidus "X" dispersés dans l'ensemble de la séquence peuvent être décrits conjointement par la clé de caractérisation "VARIANT" en employant la syntaxe "x.y" pour désigner le descripteur d'emplacement, où x et y sont les positions du premier et du dernier résidu "X", et par un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

N.B. : La représentation préférée de la séquence indiquée ci-dessus sert à fournir un listage des séquences à la date du dépôt d'une demande de brevet. La même représentation pourra ne pas être applicable à un listage des séquences fourni après cette date, car il faut tenir compte de la question de savoir si l'information fournie pourrait être prise en considération par un office de la propriété intellectuelle pour ajouter des éléments à la divulgation originale.

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.c), 7.b) et 27

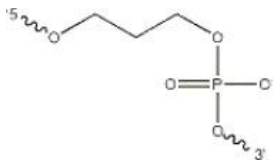
Paragraphe 3.g) – Définition de "nucléotide"

Exemple 3.g)-1 : Séquence de nucléotides interrompue par un espaceur C3

Une demande de brevet décrit la séquence ci-après :

atgcatgcatgcncggcatgcatgc

où n = un espaceur C3 dont la structure est la suivante :



Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

La séquence énumérée contient deux segments de nucléotides définis de manière spécifique et séparés par un espaceur C3.

L'espaceur C3 n'est pas un nucléotide au sens du paragraphe 3.g); le symbole conventionnel "n" est employé d'une manière non conventionnelle (voir l'Introduction au présent document). En conséquence, chaque segment est une séquence de nucléotides distincte. Étant donné que chaque segment contient plus de 10 nucléotides définis de manière spécifique, les deux segments doivent être intégrés dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Chaque segment doit être intégré dans un listage des séquences en tant que séquence distincte, chaque séquence devant disposer de son propre numéro d'identification de séquence :

atgcatgcatgc (SEQ ID NO : 5)

cggcatgcatgc (SEQ ID NO : 6)

La cytosine contenue dans chaque segment attaché à l'espaceur C3 devrait être accompagnée d'une description supplémentaire dans un tableau de caractéristiques en employant la clé de caractérisation "misc_feature" et le qualificateur "note". La valeur du qualificateur "note", qui est "free text", devrait indiquer la présence de l'espaceur, qui est lié à un autre acide nucléique, et devrait identifier l'espaceur soit par son nom chimique complet non abrégé, soit par son nom commun, par exemple espaceur C3.

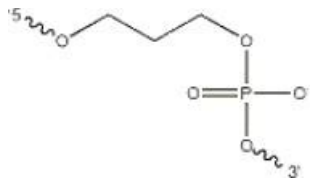
Paragraphe(s) pertinent(s) de la norme ST.26 : 3.g), 7.a) et 15

Exemple 3.g)-2 : Séquence de nucléotides avec des résidus alternatifs, y compris un espaceur C3

Une demande de brevet décrit la séquence ci-après :

atgcatgcatgcncggcatgcatgc

où n = c, a, g ou un espaceur C3 dont la structure est la suivante :



Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Il y a 24 résidus définis de manière spécifique dans la séquence énumérée interrompue par la variable "n." L'explication de la séquence figurant dans la divulgation doit être consultée pour déterminer si le "n" est employé d'une manière conventionnelle ou non conventionnelle (voir l'Introduction au présent document).

La divulgation indique n = c, a, g ou un espaceur C3. Le "n" est un symbole conventionnel employé d'une manière non conventionnelle, puisqu'il est décrit comme englobant un espaceur C3, qui ne répond pas à la définition d'un nucléotide. Le symbole "n" est également décrit comme englobant "c", "a" ou "g"; il s'ensuit que la norme ST.26 prescrit l'intégration de la séquence des 25 nucléotides dans un listage des séquences. Étant donné que deux segments séparés par l'espaceur C3 sont des séquences distinctes de la séquence des 25 nucléotides, les deux séquences de 12 nucléotides peuvent également être intégrées.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

L'exemple indique que "n = c, a, g ou un espaceur C3". Comme on l'a vu plus haut, un espaceur C3 n'est pas un nucléotide. Conformément au paragraphe 15, le symbole "n" ne doit être employé que pour représenter un nucléotide; il s'ensuit que le symbole "n" ne peut pas représenter un espaceur C3 dans un listage des séquences.

Le paragraphe 15 dit également que, lorsqu'il convient d'employer un symbole ambigu, il faudrait choisir le symbole le plus restrictif. Conformément à l'annexe I, section 1, tableau 1, le symbole "v" représente "a or c ou g", qui est plus restrictif que "n".

Lorsque, dans l'exemple, la variable "n" est c, a ou g, la séquence unique énumérée par ses résidus qui englobe les modes de réalisation les plus divulgués et est, de ce fait, la séquence la plus englobante (voir l'Introduction au présent document) à intégrer dans un listage des séquences est la suivante :

atgcatgcatgcvcggcatgcatgc (SEQ ID NO : 7)

Comme indiqué dans l'introduction au présent document, il est fortement conseillé d'intégrer toutes séquences supplémentaires de première importance pour la divulgation ou les revendications de l'invention.

Lorsque, dans l'exemple, la variable "n" est un espaceur C3, la séquence peut être considérée comme deux segments distincts de nucléotides définis de manière spécifique de chaque côté de la variable "n", c'est-à-dire atgcatgcatgc (SEQ ID NO : 8) et cggcatgcatgc (SEQ ID NO : 9). Si elles sont de première importance pour la divulgation ou les revendications, ces deux séquences devraient également être intégrées dans le listage des séquences, chacune disposant de son propre numéro d'identification de séquence.

La cytosine contenue dans chaque segment attaché à l'espaceur C3 devrait être accompagnée d'une description supplémentaire dans un tableau de caractéristiques en employant la clé de caractérisation "misc_feature" et le qualificateur "note". La valeur du qualificateur "note", qui est "free text", devrait indiquer la présence de l'espaceur, qui est lié à un autre acide nucléique, et devrait identifier l'espaceur soit par son nom chimique complet non abrégé, soit par son nom commun, par exemple espaceur C3.

N.B. : La représentation préférée de la séquence indiquée ci-dessus sert à fournir un listage des séquences à la date du dépôt d'une demande de brevet. La même représentation pourra ne pas être applicable à un listage des séquences fourni après cette date, car il faut tenir compte de la question de savoir si l'information fournie pourrait être prise en considération par un office de la propriété intellectuelle pour ajouter des éléments à la divulgation originale.

Paragraphe(s) pertinent(s) de la norme ST.26 : **3.g)**, 7.a) et 15

Exemple 3.g)-3 : Site abasique

Une demande de brevet décrit la séquence ci-après :

gagcattgac-AP-taaggct

où AP est un site abasique

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Les résidus définis d'une manière spécifique de la séquence énumérée sont interrompus par un site abasique. Le côté 5' du site abasique contient 10 nucléotides et son côté 3' en contient sept. Le paragraphe 3.g)ii)2) définit un site abasique comme un "nucléotide" lorsqu'il fait partie d'une séquence nucléotidique. Il s'ensuit que, dans cet exemple, le site abasique est considéré comme un "nucléotide" pour ce qui est de déterminer si et comment la séquence doit être intégrée dans un listage des séquences. En conséquence, les résidus se trouvant de chaque côté du site abasique font partie d'une séquence énumérée unique contenant au total 18 nucléotides, dont 17 sont définis de manière spécifique. La séquence doit donc être intégrée en tant que séquence unique comme le prescrit le paragraphe 7.a) de la norme ST.26.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence doit être intégrée dans un listage des séquences sous la forme suivante :

gagcattgacntaaggct (SEQ ID NO : 10)

Le site abasique doit être représenté par un "n" et doit être accompagné d'une description supplémentaire dans un tableau de caractéristiques. Le moyen d'annotation à préférer est la clé de caractérisation "modified_base" et le qualificateur obligatoire est le "mod_base" qui prend la valeur "OTHER". Il faut intégrer un qualificateur "note", qui décrive la base modifiée comme étant un site abasique.

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.g), 7.a) et 17

Exemple 3.g)-4 : Analogues d'acides nucléiques

Une demande de brevet divulgue la séquence d'un acide nucléique à glycol (ANG) suivante :

PO_4-tagttcattgactaaggctccccattgact-OH

où l'extrémité gauche de la séquence reproduit le côté 5' d'une séquence d'ADN.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

– Les résidus individuels qui constituent une séquence d'ANG sont considérés comme des nucléotides en vertu du paragraphe 3.g)i)2) de la norme ST.26. Il s'ensuit que la séquence compte plus de 10 nucléotides énumérés et "définis de manière spécifique" et doit être intégrée dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Les séquences d'ANG n'ont pas d'extrémité 5' ni d'extrémité 3', mais une extrémité 3' et une extrémité 2'. L'extrémité 3', qui est normalement décrite comme ayant un groupe phosphate terminal, correspond à l'extrémité 5' de l'ADN ou de l'ARN. (A noter que d'autres analogues d'acides nucléiques peuvent correspondre de façon différente à l'extrémité 5' et à l'extrémité 3' de l'ADN et de l'ARN.) En vertu du paragraphe 11, la séquence doit être intégrée dans un listage des séquences "de gauche à droite de manière à reproduire le sens 5'-3'". Elle doit donc être intégrée dans un listage des séquences sous la forme suivante :

tagttcattgactaaggctccccattgact (SEQ ID NO : 11)

La séquence doit être décrite dans le tableau de caractéristiques à l'aide de la clé de caractérisation "modified_base" et du qualificateur obligatoire "mod_base" avec l'abréviation "OTHER". Un qualificateur de type "note" doit être intégré avec le nom complet non abrégé des nucléotides modifiés, par exemple les "acides nucléiques à glycol ou les "2,3-dihydroxypropyl nuclosides". Un élément INSDFeature unique peut être employé pour décrire l'ensemble de la séquence comme un ANG où l'élément INSDFeature_location présente la série "1.30".

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.d), **3.g)**, 7.a), 11, 16, 18, 65 et 66

Paragraphe 3.k) – Définition de "défini de manière spécifique"

Exemple 3.k)-1 : Symboles ambigus des nucléotides

5' NNG KNG KNG K 3'

N et K sont des codes ambigus IUPAC-IUB

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

NON

Les codes ambigus IUPAC-IUB correspondent à la liste des symboles des nucléotides définis à l'annexe I, section 1, tableau 1. Selon le paragraphe 3.k), un nucléotide défini de manière spécifique est un nucléotide différent de ceux qui sont représentés par le symbole "n" et énumérés à l'annexe I. En conséquence, "K" et "G" sont des nucléotides définis de manière spécifique tandis que "N" n'en est pas un.

La séquence énumérée n'ayant pas au moins 10 nucléotides définis de manière spécifique, le paragraphe 7.a) de la norme ST.26 ne prescrit pas son intégration dans un listage des séquences.

Question 2 : La norme ST.26 autorise-t-elle l'intégration de la ou des séquences?

NON

En vertu du paragraphe 8, "un listage des séquences ne doit contenir aucune séquence comportant moins de 10 nucléotides définis de manière spécifique…". N'ayant pas au moins 10 nucléotides définis de manière spécifique, la séquence énumérée ne doit pas être intégrée dans un listage des séquences.

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.k), 7.a), 8 et 13

Exemple 3.k)-2 : Symbole ambigu "n" employé d'une manière à la fois conventionnelle et non conventionnelle

Une demande divulgue la séquence artificielle suivante : 5'-AATGCCGGAN-3'. La divulgation indique également ce qui suit :

i) dans un mode de réalisation, N représente tout nucléotide;

ii) dans un mode de réalisation, N est facultatif mais représente de préférence G;

iii) dans un mode de réalisation, N représente K;

iv) dans un mode de réalisation, N représente C.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

NON

La séquence énumérée contient neuf nucléotides définis de manière spécifique et un "N." Il faut consulter l'explication de la séquence donnée dans la divulgation afin de déterminer si le symbole "N" est utilisé d'une manière conventionnelle (voir l'Introduction au présent document).

L'examen des modes de réalisation divulgués i) à iv) de la séquence énumérée montre que le mode de réalisation de "N" le plus englobant est "tout nucléotide". Dans le mode de réalisation le plus englobant, "N" est utilisé d'une manière conventionnelle dans la séquence énumérée.

Pour certains modes de réalisation, "N" est décrit comme représentant des résidus définis de manière spécifique (c'est-à-dire "N représente C" dans la partie iv)). Toutefois, seul le mode de réalisation le plus englobant (c'est-à-dire "N représente tout nucléotide") est pris en considération lorsqu'il s'agit de déterminer si une séquence doit être intégrée dans un listage des séquences. Il s'ensuit que la séquence énumérée qui doit être évaluée est 5'-AATGCCGGAN-3'.

Sur la base de cette analyse, la séquence énumérée, à savoir AATGCCGGAN, ne contient pas 10 nucléotides définis de manière spécifique. De ce fait, le paragraphe 7.a) de la norme ST.26 ne prescrit pas l'intégration de la séquence dans un listage des séquences, en dépit du fait que "n" est également défini comme des nucléotides particuliers dans certains modes de réalisation.

Question 2 : La norme ST.26 autorise-t-elle l'intégration de la ou des séquences?

NON

La séquence "AATGCCGGAN" ne doit pas être intégrée dans un listage des séquences.

Toutefois, une séquence alternative décrite peut être intégrée dans un listage des séquences si le "N" est remplacé par un nucléotide défini de manière spécifique.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

L'intégration des séquences représentant des modes de réalisation qui constituent une partie essentielle de l'invention est fortement conseillée. L'intégration de ces séquences permet d'effectuer une recherche plus approfondie et porte à la connaissance du public l'objet pour lequel la délivrance d'un brevet est demandée.

En ce qui concerne l'exemple ci-dessus, il est vivement recommandé d'intégrer dans le listage des séquences les trois séquences supplémentaires ci-après, chacune disposant de son propre numéro d'identification de séquence :

aatgccggag (SEQ ID NO : 12)

aatgccggak (SEQ ID NO : 13)

aatgccggac (SEQ ID NO : 14)

Si les séquences susmentionnées ne sont pas intégrées toutes les trois, le nucléotide qui remplace le "n" devrait être annoté pour décrire les alternatives. Par exemple, si seule la séquence SEQ ID NO : 12 ci-dessus est intégrée dans le listage des séquences, la clé de caractérisation "misc_difference" avec l'emplacement de caractéristique "10" doit être employée avec deux qualificateurs du type "replace" lorsque la valeur de l'un serait "k" et la valeur de l'autre serait "c".

N.B. : La représentation préférée de la séquence indiquée ci-dessus sert à fournir un listage des séquences à la date du dépôt d'une demande de brevet. La même représentation pourra ne pas être applicable à un listage des séquences fourni après cette date, car il faut tenir compte de la question de savoir si l'information fournie pourrait être prise en considération par un office de la propriété intellectuelle pour ajouter des éléments à la divulgation originale.

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.k), 7.a), 8 et 13

Exemple 3.k)-3 : Symbole ambigu "n" employé d'une manière non conventionnelle

Une demande divulgue la séquence : 5'-aatgttggan-3'

où n représente c

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Conformément au paragraphe 3.k), un nucléotide "défini de manière spécifique" est tout nucléotide différent de ceux qui sont représentés par le symbole "n" énumérés dans l'annexe I, section 1, tableau 1.

Dans cet exemple, "n" est employé d'une manière non conventionnelle pour représenter uniquement "c". La divulgation n'indique pas que "n" est employé d'une manière conventionnelle pour représenter "tout nucléotide". La séquence doit donc être interprétée comme si le symbole conventionnel équivalent, c'est-à-dire "c", avait été employé dans la séquence (voir l'Introduction au présent document). Il s'ensuit que la séquence énumérée à considérer est la suivante :

5'-aatgttggac-3'

Cette séquence contient 10 nucléotides définis de manière spécifique et doit, conformément au paragraphe 7.a) de la norme ST.26, être intégrée dans le listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence doit être intégrée dans un listage des séquences sous la forme suivante : aatgttggac (SEQ ID NO : 15)

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.k) et 7.a)

Exemple 3.k)-4 : Les symboles ambigus autres que "n" sont "définis de manière spécifique

Une demande de brevet décrit la séquence ci-après :

5' NNG KNG KNG KAG VCR 3'

où N, K, V et R sont des codes ambigus IUPAC-IUB

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Les codes ambigus IUPAC-IUB correspondent à la liste des symboles de nucléotides définis dans l'annexe I, section 1, tableau 1. Conformément au paragraphe 3.k), un nucléotide "défini de manière spécifique" est tout nucléotide différent de ceux qui sont représentés par le symbole "n" énumérés dans l'annexe I, section 1, tableau 1. "K", "V" et "R" sont donc des nucléotides "définis de manière spécifique".

Cette séquence contient 11 nucléotides énumérés et "définis de manière spécifique" et doit, conformément au paragraphe 7.a) de la norme ST.26, être intégrée dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence doit être intégrée dans un listage des séquences sous la forme suivante :

nngkngkngkagvcr (SEQ ID NO : 16)

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.k), 7.a) et 15

Exemple 3.k)-5 : Abréviation ambiguë "Xaa" employée d'une manière non conventionnelle

Une demande de brevet décrit la séquence ci-après :

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

où Xaa à la position 1 représente tout acide aminé, Xaa à la position 4 représente Lys, Xaa à la position 5 représente Gly et Xaa à la position 6 représente la leucine ou l'isoleucine.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Le peptide énuméré dans la formule contient trois acides aminés définis de manière spécifique aux positions 2, 3 et 7. Le premier acide aminé est représenté par une abréviation conventionnelle, à savoir Xaa, qui représente tout acide aminé. Toutefois, les 4e, 5e et 6e acides aminés sont représentés par une abréviation conventionnelle employée d'une manière non conventionnelle (voir l'Introduction au présent document). De ce fait, l'explication de la séquence donnée dans la divulgation est consultée pour déterminer la définition de "Xaa" à ces positions. Étant donné que les "Xaa" aux positions 4 à 6 désignent un acide aminé particulier, la séquence doit être interprétée comme si les abréviations conventionnelles équivalentes avaient été employées dans la séquence, à savoir Lys, Gly et (Leu or Ile). Il s'ensuit que la séquence contient au moins quatre acides aminés définis de manière spécifique et doit être intégrée dans un listage des séquences comme le prescrit le paragraphe 7.b) de la norme ST.26.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence emploie une abréviation conventionnelle "Xaa" d'une manière non conventionnelle. De ce fait, l'explication de la séquence donnée dans la divulgation est consultée pour déterminer la définition de "Xaa" aux positions 4, 5 et 6. L'explication définit "Xaa" comme une lysine à la position 4, une glycine à la position 5 et une leucine ou isoleucine à la position 6. Les symboles conventionnels pour ces acides aminés sont K, G et J, respectivement. En conséquence, la séquence devrait être représentée dans le listage des séquences sous la forme suivante :

XYEKGJL (SEQ ID NO : 17)

Conformément au paragraphe 27, "X" sera considéré comme l'équivalent de 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 dans le tableau de caractéristiques. Étant donné que "X" à la position 1 de la séquence SEQ ID NO : 17 représente "tout acide aminé", il doit être annoté au moyen de la clé de caractérisation "VARIANT" et d'un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

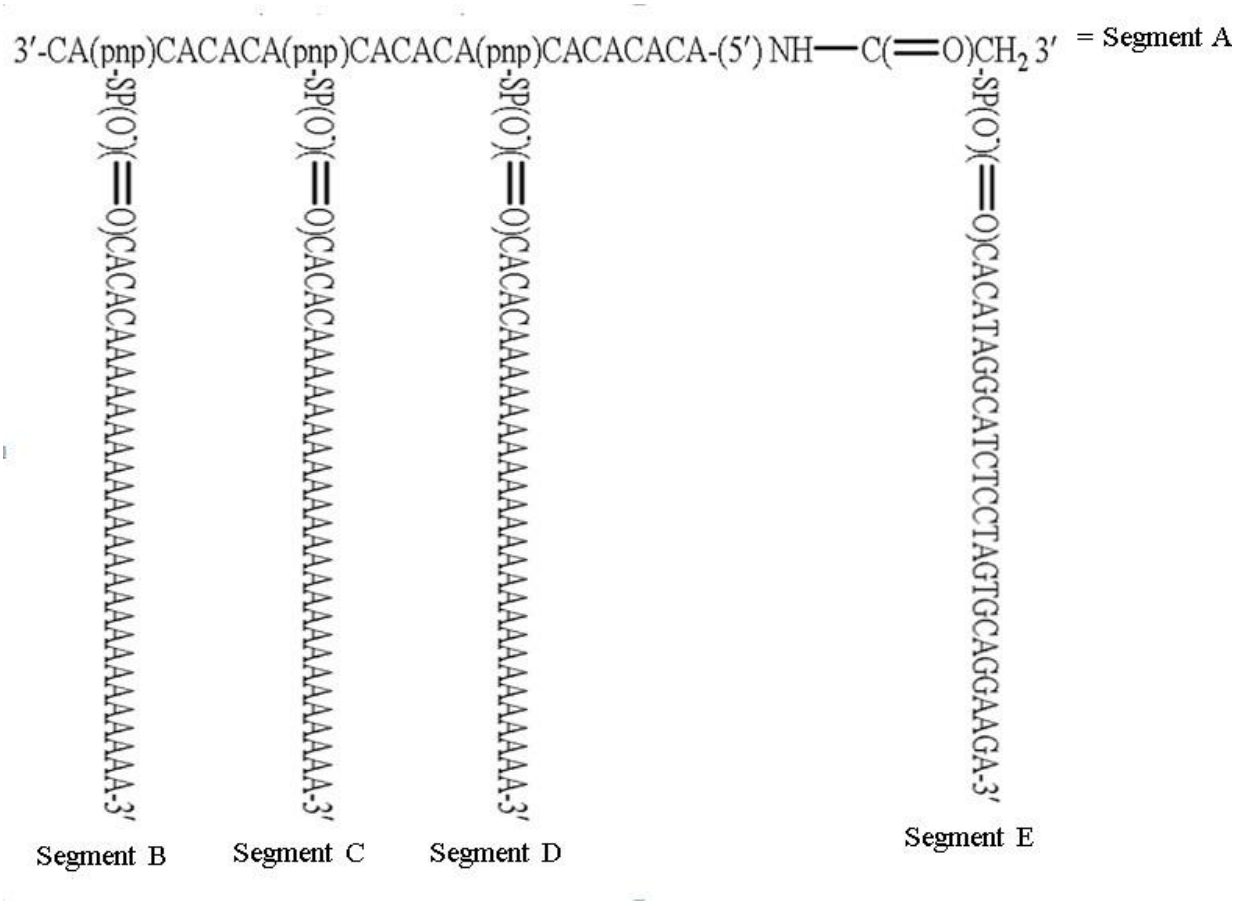
Chaque fois que possible, chaque "X" devrait être annoté individuellement. Cependant, une région de résidus "X" contigus ou un grand nombre de résidus "X" dispersés dans l'ensemble de la séquence peuvent être décrits conjointement par la clé de caractérisation "VARIANT" en employant la syntaxe "x.y" pour désigner le descripteur d'emplacement, où x et y sont les positions du premier et du dernier résidu "X", et par un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.k), 7.b), 26 et 27

Paragraphe 7.a) – Séquences de nucléotides à intégrer dans un listage

Exemple 7.a)-1 : Séquence de nucléotides ramifiée

La description divulgue la séquence de nucléotides ramifiée ci-après :



où "pnp" est une liaison ou un monomère contenant une fonctionnalité bromoacetylamino;

3'-CA(pnp)CACACA(pnp)CACACA(pnp)CACACACA-(5')NH—C(=O)CH2 3' est le segment A;

SP(O-)(=O)CACACAAAAAAAAAAAAAAAAAAAA 3' représente les segments B, C et D;

et SP(O-)(=O)CACATAGGCATCTCCTAGTGCAGGAAGA 3' est le segment E.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI – les quatre segments verticaux B à E doivent être intégrés dans un listage des séquences.

NON – le segment horizontal A ne doit pas être intégré dans un listage des séquences.

La figure ci-dessus est un exemple de séquence d'acides nucléiques ramifiée "en peigne" contenant cinq segments linéaires : le segment horizontal A et les quatre segments verticaux B à E.

Conformément au paragraphe 7.a), les régions linéaires de séquences de nucléotides ramifiées contenant au moins 10 nucléotides définis de manière spécifique, et dont les nucléotides adjacents sont reliés par une liaison de 3' à 5', doivent être intégrées dans un listage des séquences.

Les quatre segments verticaux B à E contiennent chacun plus de 10 nucléotides définis de manière spécifique, et dont les nucléotides adjacents sont reliés par une liaison 3' à 5'; ils doivent donc tous être intégrés dans un listage des séquences.

S'agissant du segment horizontal A, les régions linéaires de la séquence de nucléotides sont reliées par le

fragment non nucléotidique "pnp" et chacune de ces portions linéaires reliées contient moins de 10 nucléotides définis de manière spécifique. De ce fait, étant donné qu'aucune région du segment A ne contient au moins 10 nucléotides définis de manière spécifique, et dont les nucléotides adjacents soient reliés par une liaison de 3' à 5', le paragraphe 7.a) de la norme ST.26 ne prescrit pas leur intégration dans le listage des séquences.

Question 2 : La norme ST.26 autorise-t-elle l'intégration de la ou des séquences?

NON

Conformément au paragraphe 8, "un listage des séquences ne doit contenir aucune séquence comportant moins de 10 nucléotides définis de manière spécifique…"

Aucune région du segment A ne contient au moins 10 nucléotides définis de manière spécifique et dont les nucléotides adjacents soient reliés par une liaison de 3' à 5'; en conséquence, elle ne doit pas être intégrée dans un listage des séquences en tant que séquence distincte disposant de son propre numéro d'identification de séquence.

Toutefois, les segments B, C, D et E peuvent faire l'objet d'une annotation indiquant qu'ils sont reliés au segment A.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Les segments B, C et D sont identiques et doivent être intégrés dans un listage des séquences en tant que séquence unique :

cacacaaaaaaaaaaaaaaaaaaaaaaaaa (SEQ ID NO : 18)

Le premier "c" de la séquence devrait s'accompagner d'une description supplémentaire à l'aide de la clé de caractérisation "misc_feature" et du qualificateur de type "note" prenant une valeur qui serait par exemple la suivante : "Cette séquence est l'un des quatre rameaux d'un polynucléotide ramifié."

Le segment E doit être intégré dans un listage des séquences en tant que séquence

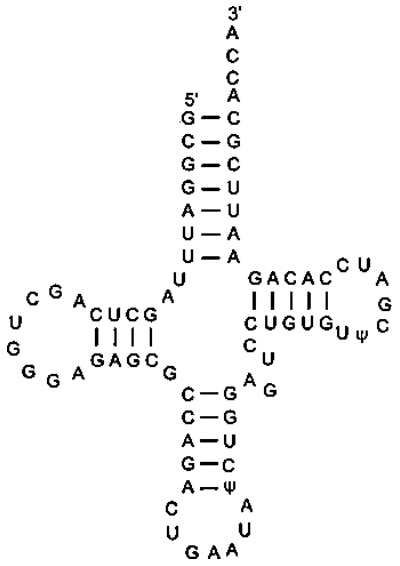
unique : cacataggcatctcctagtgcaggaaga (SEQ ID NO : 19)

Le premier "c" de la séquence devrait s'accompagner d'une description supplémentaire à l'aide de la clé de caractérisation "misc_feature" et du qualificateur de type "note" prenant une valeur qui serait par exemple la suivante : "Cette séquence est l'un des quatre rameaux d'un polynucléotide ramifié."

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.a), 8, 11, 13 et 17

Exemple 7.a)-2 : Séquence de nucléotides linéaire comportant une structure secondaire

Une demande de brevet décrit la séquence ci-après :



où Ψ est la pseudouridine.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

La séquence nucléotidique contient 73 nucléotides énumérés et définis de manière spécifique. L'exemple présente donc au moins 10 nucléotides "définis de manière spécifique" et, comme le prescrit le paragraphe 7.a) de la norme ST.26, doit être intégré dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La consultation de la divulgation indique que "Ψ" est l'équivalent de la pseudouridine. Le seul symbole conventionnel qui peut servir à représenter la pseudouridine est "n"; il s'ensuit que le "Ψ" est un symbole non conventionnel employé pour représenter le symbole conventionnel "n" (voir l'Introduction au présent document). En conséquence, la séquence doit être interprétée comme ayant deux symboles "n" à la place des deux symboles "Ψ".

Le symbole "u" ne doit pas être employé pour représenter l'uracile dans une molécule d'ARN dans le listage des séquences. Conformément au paragraphe 14, le symbole "t" désigne l'uracile dans de l'ARN. La séquence doit être intégrée sous la forme suivante :

gcggatttagctcagctgggagagcgccagactgaatanctggagtcctgtgtncgatccacagaattcgcacca (SEQ ID NO : 20)

La valeur du qualificateur obligatoire "mol_type" de la clé de caractérisation "source" est "tRNA". Des informations supplémentaires peuvent être fournies avec la clé de caractérisation "tRNA" et un ou des qualificateurs appropriés.

Les résidus "n" doivent s'accompagner d'une description supplémentaire dans un tableau de caractéristiques en employant la clé de caractérisation "modified_base" et le qualificateur obligatoire "mod_base" avec l'abréviation "p" pour la pseudouridine en tant que valeur du qualificateur (voir l'annexe 1, tableau 2).

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.a), 11, 13, 14, 17, 62, 84 et annexe I, sections 2 et 5, clé de caractérisation 5.43

Exemple 7.a)-3 : Symboles ambigus des nucléotides employés d'une manière non conventionnelle

Une demande de brevet décrit la séquence ci-après :

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

L'explication de la séquence donnée dans la divulgation indique par ailleurs ce qui suit : "Un "DR Element" est constitué par la séquence 5' ATCAGCCAT 3'. Un DR Element mutant, ou MDR, est un élément DR dont les cinq nucléotides du milieu, CAGCC, deviennent après mutation TTTTT."

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

La séquence énumérée emploie le symbole "MDR". Lorsque l'on ignore si un symbole employé dans une séquence est censé être un symbole conventionnel, c'est-à-dire un symbole indiqué dans l'annexe 1, section 3, tableau 3, ou un symbole non conventionnel, l'explication de la séquence donnée dans la divulgation doit être consultée pour le déterminer (voir l'Introduction au présent document). Conformément au tableau 3, "MDR" pourrait être interprété comme désignant trois symboles conventionnels (m = a ou c, d = a ou g ou t/u, r = g ou a) ou une abréviation qui est une notation topologique pour une autre structure.

La consultation de la divulgation indique qu'un élément MDR est l'équivalent de 5' ATTTTTTAT 3'. Les lettres "MDR" sont considérées comme des symboles conventionnels employés d'une manière non conventionnelle; la séquence doit donc être interprétée comme si elle était divulguée à l'aide des symboles conventionnels équivalents. En conséquence, la séquence énumérée que l'on envisage d'intégrer dans un listage des séquences est la suivante :

5' GATC ATTTTTTAT ATTTTTTAT ATTTTTTAT ATTTTTTAT GTAC 3'

La séquence énumérée comptant 44 nucléotides définis d'une manière spécifique, elle doit, conformément au paragraphe 7.a) de la norme ST.26, être intégrée dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence doit être intégrée dans un listage des séquences sous la forme suivante :

gatcatttttatatttttatatttttatatttttatgtac (SEQ ID NO : 21)

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.a) et 13

Exemple 7.a)-4 : Symboles ambigus des nucléotides employés d'une manière non conventionnelle

Une demande de brevet décrit la séquence ci-après :

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

L'explication de la séquence donnée dans la divulgation indique par ailleurs que "N" est constitué par la séquence 5' ATACGCACT 3'.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

La séquence énumérée emploie le symbole "N". L'explication de la séquence donnée dans la divulgation doit être consultée pour déterminer si le "N" est employé d'une manière conventionnelle ou non conventionnelle (voir l'Introduction au présent document).

La consultation de la divulgation indique que "N" est l'équivalent de 5' ATACGCACT 3'. Le "N" est donc un symbole conventionnel employé d'une manière non conventionnelle. En conséquence, la séquence doit être interprétée comme si elle était divulguée à l'aide des symboles conventionnels équivalents :

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

La séquence énumérée comptant 44 nucléotides définis d'une manière spécifique, elle doit, conformément au paragraphe 7.a) de la norme ST.26, être intégrée dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence doit être intégrée dans un listage des séquences sous la forme suivante :

attcatacgcactatacgcactatacgcactatacgcactgtac (SEQ ID NO : 22)

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.a) et 13

Exemple 7.a)-5 : Symboles des nucléotides non conventionnels

Une demande de brevet décrit la séquence ci-après :

5' GATC-β-β-β-β-GTAC 3'

L'explication de la séquence donnée dans la divulgation indique par ailleurs que "β" est constitué par la séquence 5' ATACGCACT 3'.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

La séquence énumérée emploie le symbole non conventionnel "β". L'explication de la séquence donnée dans la divulgation doit être consultée pour déterminer la signification de "β" (voir l'Introduction au présent document).

La consultation de la divulgation indique que "β" est l'équivalent de 5' ATACGCACT 3'. Le "β" est donc un symbole non conventionnel employé pour représenter une séquence de neuf symboles conventionnels définis d'une manière spécifique. En conséquence, la séquence doit être interprétée comme si elle était divulguée à l'aide des symboles conventionnels équivalents :

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

La séquence énumérée comptant 44 nucléotides définis d'une manière spécifique, elle doit, conformément au paragraphe 7.a) de la norme ST.26, être intégrée dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence doit être intégrée dans un listage des séquences sous la forme suivante :

gatcatacgcactatacgcactatacgcactatacgcactgtac (SEQ ID NO : 23)

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.a) et 13

Exemple 7.a)-6 : Symboles de nucléotides non conventionnels

Une demande de brevet décrit la séquence ci-après :

5' GATC-β-β-β-β-GTAC 3'

L'explication de la séquence donnée dans la divulgation indique par ailleurs que "β" est égal à l'adénine, à l'inosine ou à la pseudouridine.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

NON

La séquence énumérée emploie le symbole non conventionnel "β". L'explication de la séquence donnée dans la divulgation doit être consultée pour déterminer la signification de "β" (voir l'Introduction au présent document).

La consultation de la divulgation indique que "β" est l'équivalent de l'adénine, de l'inosine ou de la pseudouridine. Le seul symbole conventionnel qui peut être employé pour représenter "adénine, inosine ou pseudouridine" est "n"; le "β" est donc un symbole non conventionnel employé pour représenter le symbole conventionnel "n". En conséquence, la séquence doit être interprétée comme ayant quatre symboles "n" (indiqué par "N" ci-dessous) à la place des quatre symboles "β" :

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

La séquence énumérée ne comportant que huit nucléotides définis de manière spécifique, elle ne doit pas, conformément au paragraphe 7.a) de la norme ST.26, être intégrée dans un listage des séquences.

Question 2 : La norme ST.26 autorise-t-elle l'intégration de la ou des séquences?

NON

La séquence énumérée, 5' GATC-N-N-N-N-GTAC 3', ne doit pas être intégrée dans un listage des séquences.

Toutefois, une séquence alternative divulguée peut être intégrée dans un listage des séquences si au moins deux des symboles "n" sont remplacés par l'adénine, ce qui donnerait une séquence d'au moins 10 nucléotides définis de manière spécifique.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Une représentation autorisée possible est la suivante :

gatcaaaagtac (SEQ ID NO : 24)

Dans l'exemple ci-dessus, les quatre nucléotides d'adénine qui remplacent les symboles β devraient s'accompagner d'une annotation indiquant que ces positions pourraient être remplacées par l'inosine ou la pseudouridine.

La clé de caractérisation "misc_difference" devrait être employée avec un emplacement de caractéristique 5-8 et un qualificateur du type "note" prenant la valeur "Un nucléotide aux positions 5-8 peut être remplacé par l'inosine ou la pseudouridine", par exemple. Étant donné que ces alternatives sont des nucléotides modifiés, la clé de caractérisation "modified_base" et le qualificateur "mod_base" seraient requis. La valeur du qualificateur "mod_base" peut être "OTHER" avec un qualificateur du type "note" et la valeur de "i ou p".

D'autres permutations sont possibles.

N.B. : La représentation préférée de la séquence indiquée ci-dessus sert à fournir un listage des séquences à la date du dépôt d'une demande de brevet. La même représentation pourra ne pas être applicable à un listage des séquences fourni après cette date, car il faut tenir compte de la question de savoir si l'information fournie pourrait être prise en considération par un office de la propriété intellectuelle pour ajouter des éléments à la divulgation originale.

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.a), 8, 13 et 17

Exemple 7.a)-7 : Nucléotides inversés I

Une demande de brevet divulgue la séquence d'ADN simple brin suivante :

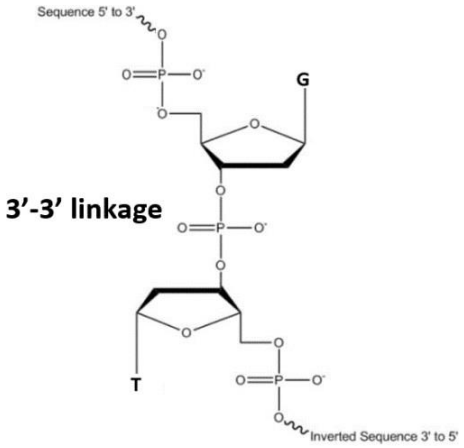
5 ' attgactaagtg ttccccattgact 5 '

Dans cette séquence, la direction de la séquence change à l'intérieur du brin en raison d'une liaison inversée 3' – 3' entre les résidus 12 et 13. La partie soulignée de la séquence est orientée dans le sens 3' – 5' de gauche à droite.

Question 1 : La norme ST.26 prescrit-elle l'inclusion de la ou des séquences?

OUI

Les 12 premiers résidus sont représentés dans l'orientation standard 5' – 3'. Le "g" en position 12 est lié au "t" en position 13 par une liaison inversée 3' – 3'. Le "g" en position 12 est lié au "t" en position 13 par une liaison inversée 3' – 3' :



Le reste de la molécule, représenté dans les positions 13 - 25, est dans l'orientation opposée – 3' – 5'. Le paragraphe 11 de la norme ST.26 exige qu'une séquence de nucléotides soit représentée dans le sens 5' – 3', de gauche à droite.

Par conséquent, pour représenter correctement cette molécule dans le listage des séquences, elle doit être représentée par deux séquences - une séquence pour les positions 1 - 12, et une seconde pour les positions 13 - 25. Chaque partie de la séquence contient au moins dix nucléotides "spécifiquement définis" et, conformément au paragraphe 7.a) de la norme ST.26, doit être incluse dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Les positions 1-12 doivent être intégrées dans un listage des sequences sous la forme suivante :

attgactaagtg (SEQ ID NO: 99)

La position 12 doit être décrite dans un tableau de caractéristiques comportant la clé de caractérisation "misc_feature" et le qualificateur "note" avec une valeur indiquant que le résidu est connecté à une séquence nucléotidique inversée par une liaison phosphodiester 3' – 3' avec un monophosphate de thymidine 3'.

Les positions 13-25 doivent intégrées dans un listage des sequences sous la forme suivante :

tcagttacccctt (SEQ ID NO: 100)

Il convient de noter que cette séquence est inversée par rapport à la manière dont elle a été décrite dans la divulgation originale, c'est-à-dire qu'elle est maintenant orientée dans le sens 5' – 3', de gauche à droite. La position 13 doit être décrite dans un tableau de caractéristiques comportant la clé de caractérisation "misc_feature" et le qualificateur "note" avec une valeur indiquant que le résidu est connecté à une séquence de nucléotides inversée par une liaison phosphodiester 3'-3' avec une guanosine 3'-monophosphate.

Paragraphes pertinents de la norme ST.26 : 7.a), 11

Exemple 7.a)-8 : Nucléotides inversés II

Une demande de brevet divulgue la séquence d'ADN suivante :

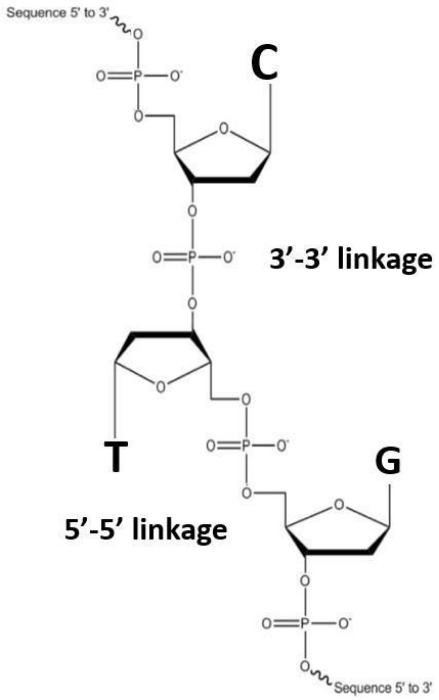
5 ' - a t t g a c t a a g t g c **t** g c c c a t t g a c t - 3 '

Dans cette sequence, le résidu thymine souligné (position 14) est un nucléotide inversé qui est relié à la cytosine par une liaison phosphodiester 3' – 3' et à la guanine par une liaison phosphodiester 5' - 5'.

Question 1 : La norme ST.26 prescrit-elle l'inclusion de la ou des séquences?

OUI

La thymidine inversée en position 14 interrompt la direction 5' - 3' de la séquence en introduisant une liaison 3' – 3' entre les résidus 13 et 14 et une liaison 5' – 5' entre les résidus 14 et 15 :



La thymidine inversée relie la première partie de la séquence, les résidus 1 - 13, et la deuxième partie de la séquence, les résidus 15 - 25. Chaque partie de la séquence contient au moins 10 nucléotides énumérés et définis de manière spécifique. Par conséquent, chaque partie de la séquence doit être incluse dans un listage des séquences comme l'exige le paragraphe 7.a) de la norme ST.26. La thymidine unique qui relie les deux parties de la séquence ne peut pas être incluse dans le listage des séquences en tant que séquence distincte parce qu'il ne s'agit pas d'une séquence d'au moins 10 nucléotides définis de manière spécifique.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La thymidine inversée en position 14 interrompt la direction 5' – 3' de la séquence en introduisant une liaison 3' – 3' entre les résidus 13 et 14 et une liaison 5' – 5' entre les résidus 14 et 15. Le paragraphe 11 de la norme ST.26 exige qu'une séquence de nucléotides soit représentée dans le sens 5' à 3', de gauche à droite. Par conséquent, pour représenter correctement cette molécule dans le listage des séquences, elle doit être représentée par deux séquences - une première séquence pour les résidus 1 - 13 et une seconde séquence pour les résidus 15 - 25.

Les positions 1 - 13 doivent être intégrées dans un listage des sequences sous la forme suivante :

attgactaagtgc (SEQ ID NO: 101)

La position 13 doit être décrite dans une tableau de caractéristiques comportant la clé de caractérisation "misc_feature" et le qualificateur "note" avec une valeur indiquant que le résidu est relié à une autre séquence par une liaison phosphodiester 3' – 3' à une thymidine 3'-monophosphate qui est à son tour reliée à une autre séquence par une liaison phosphodiester 5' – 5'.

Les positions 15-25 doivent être intégrées dans un listage des sequences sous la forme suivante :

gcccattgact (SEQ ID NO: 102)

La position 1 doit être décrite dans une tableau de caractéristiques comportant la clé de caractérisation "misc_feature" et le qualificateur "note" avec une valeur indiquant que le résidu est relié à une autre séquence par une liaison phosphodiester 5' – 5' à une thymidine 3'-monophosphate qui est à son tour reliée à une autre séquence par une liaison phosphodiester 3' – 3'.

Paragraphes pertinents de la norme ST.26 : 7.a), 11

Paragraphe 7.b) – Séquences d'acides aminés à intégrer dans un listage

Exemple 7.b)-1 : Au moins quatre acides aminés définis de manière spécifique

XXXXXXXXDXXXXXXXXXXFXXXXXXXXXXXXXXXXXXXXXXXXXXXXXAXXXXXXXXXXXXXXXXXXXXGXXXX
X

où X = tout acide aminé

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Le peptide énuméré contient quatre acides aminés définis de manière spécifique. Le symbole "X" est employé de manière conventionnelle pour représenter les autres acides aminés comme étant tout acide aminé (voir l'Introduction au présent document).

Étant donné qu'il y a quatre acides aminés définis de manière spécifique, à savoir Asp, Phe, Ala et Gly, le paragraphe 7.b) de la norme ST.26 prescrit l'intégration de la séquence dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence doit être représentée sous la forme suivante :

XXXXXXXXDXXXXXXXXXXFXXXXXXXXXXXXXXXXXXXXXXXXXXXXXAXXXXXXXXXXXXXXXXXXXXGXXXX
X (SEQ ID NO : 25)

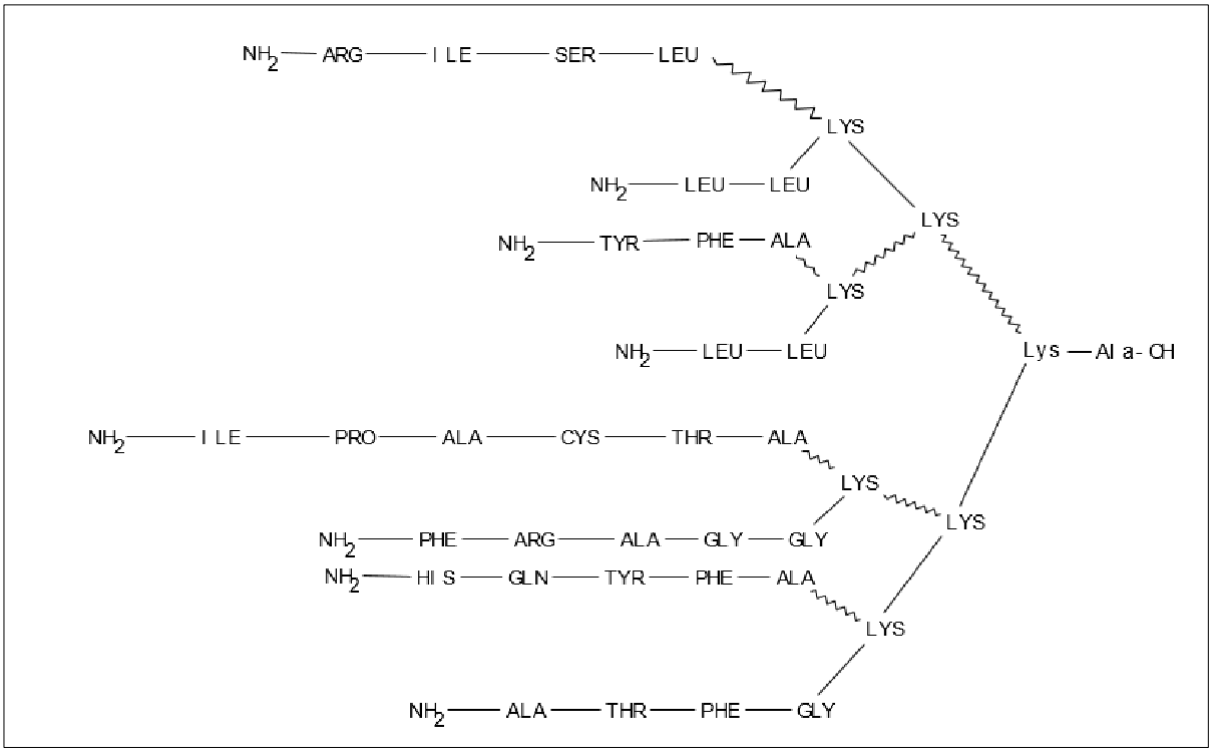
Conformément au paragraphe 27, "X" sera considéré comme l'équivalent de l'un des symboles "A", "R", "N", "D", "C", "Q", "E", "G", "H", "I", "L", "K", "M", "F", "P", "O", "S", "U", "T", "W", "Y" ou "V", sauf s'il est accompagné d'une description supplémentaire dans le tableau de caractéristiques. Comme "X" dans la séquence SEQ ID NO : 25 représente "tout acide aminé", il doit être annoté au moyen de la clé de caractérisation "VARIANT" et d'un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

Chaque fois que possible, chaque "X" devrait être annoté individuellement. Cependant, une région de résidus "X" contigus ou un grand nombre de résidus "X" dispersés dans l'ensemble de la séquence peuvent être décrits conjointement par la clé de caractérisation "VARIANT" en employant la syntaxe "x.y" pour désigner le descripteur d'emplacement, où x et y sont les positions du premier et du dernier résidu "X", et par un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.b), 8 et 27

Exemple 7.b)-2 : Séquence d'acides aminés ramifiée

La demande décrit une séquence ramifiée dans laquelle les résidus de lysine sont utilisés comme squelette pour former huit rameaux auxquels sont reliées plusieurs chaînes peptidiques linéaires. La lysine est un acide aminé dibasique, qui dispose de deux sites de liaison peptidique. Le peptide est illustré comme suit :

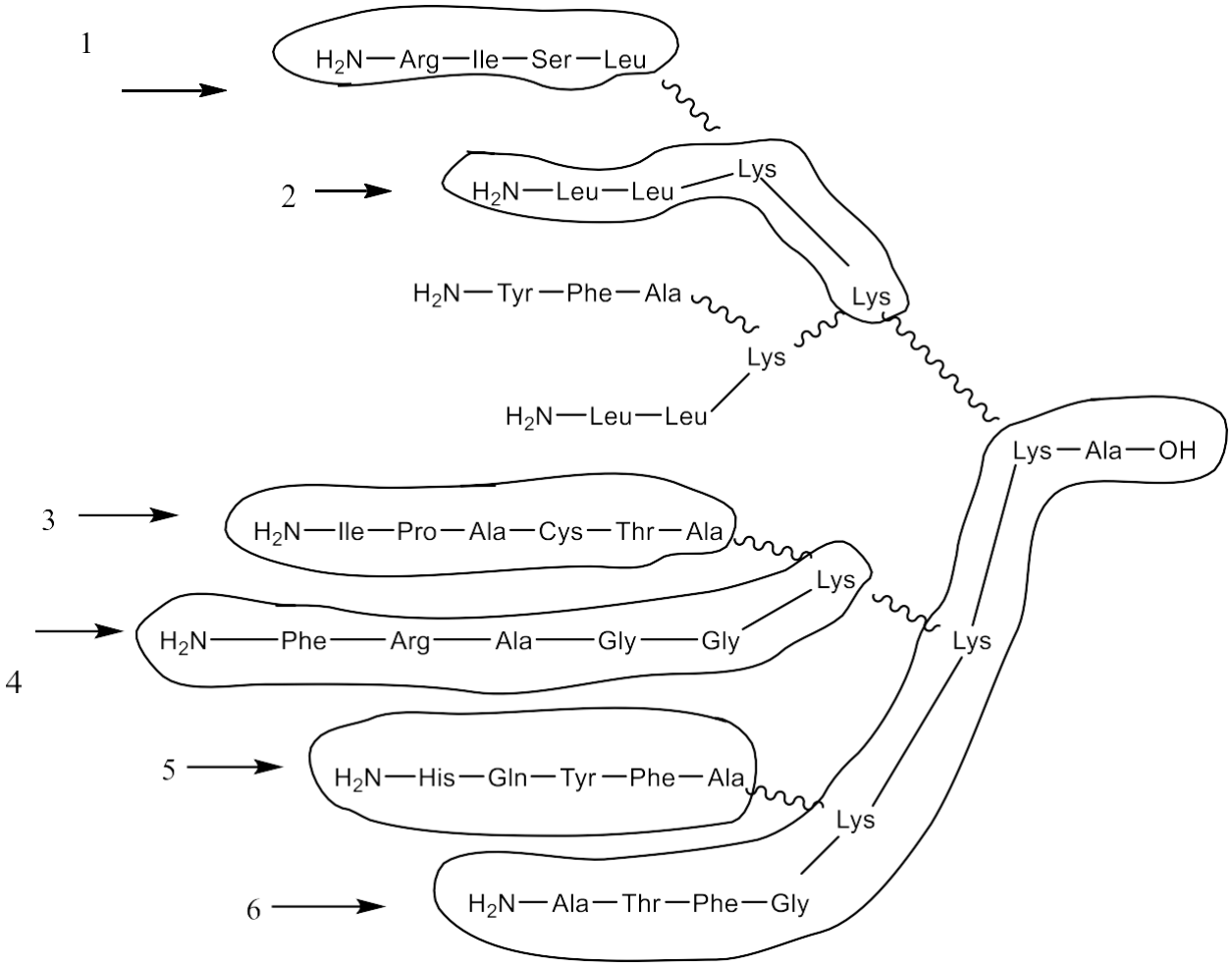


Dans le peptide ramifié ci-dessus, les liens entre la lysine et un autre acide aminé décrits par représentent une liaison amide entre la terminaison amine de la lysine et la terminaison carboxyle de l'acide aminé lié. Les liens décrits par représentent une liaison amide entre la chaîne latérale de la lysine et la terminaison carboxyle de l'acide aminé lié.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

L'exemple divulgue une séquence ramifiée où les résidus de lysine sont utilisés comme un squelette. Le paragraphe 7.b) prescrit l'intégration dans un listage des séquences de la région non ramifiée ou linéaire de la séquence, qui contient au moins quatre acides aminés définis de manière spécifique. Dans l'exemple ci-dessus, les régions linéaires du peptide ramifié qui comportent au moins quatre acides aminés définis de manière spécifique sont cerclées :



Le paragraphe 7.b) de la norme ST.26 prescrit l'intégration des peptides 1 à 6 ci-dessus dans un listage des séquences.

Les peptides dont l'intégration dans un listage des séquences n'est pas prescrite sont les suivants :

YFA

LLK

Question 2 : La norme ST.26 autorise-t-elle l'intégration de la ou des séquences?

NON

Conformément au paragraphe 8, "un listage des séquences ne doit contenir aucune séquence comportant moins de quatre acides aminés définis de manière spécifique

Les peptides YFA et LLK ne contiennent chacun que trois acides aminés définis de manière spécifique et ne doivent donc pas être inclus dans un listage des séquences en tant que séquences distinctes disposant de leur propre numéro d'identification de séquence.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Les peptides 1 à 6 doivent être représentés par des identificateurs de séquence distincts :

RISL (SEQ ID NO : 26)

LLKK (SEQ ID NO : 27)

IPACTA (SEQ ID NO : 28)

FRAGGK (SEQ ID NO : 29)

HQYFA (SEQ ID NO : 30)

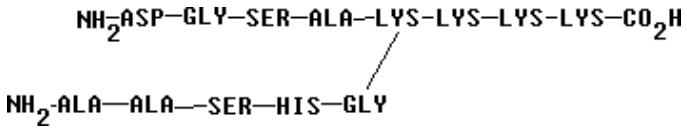
ATFGKKKA (SEQ ID NO : 31)

La structure ramifiée peut être annotée à l'aide de la clé de caractérisation "SITE" et du qualificateur obligatoire "note", qui peut par exemple prendre la valeur "This sequence is one part of a branched amino acid sequence" ("Cette séquence fait partie d'une séquence d'acides aminés ramifiée"). Selon le paragraphe 30 de la norme ST.26, les numéros d'identification de séquence 27, 29 et 31 doivent comporter pour chaque lysine une annotation indiquant qu'il s'agit d'un acide aminé modifié; on utilisera la clé de caractérisation "SITE" ainsi que le qualificateur "note" pour indiquer que la chaîne latérale de la lysine est reliée à une autre séquence par une liaison amide. Les numéros d'identification de séquence 26, 28 et 30 devraient s'accompagner d'une annotation indiquant que le terminal C d'un acide aminé est lié à une autre séquence; on utilisera à cette fin la clé de caractérisation "SITE" ainsi que le qualificateur "note".

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.b), 8, 26, 29, 30 et 31

Exemple 7.b)-3 : Séquence d'acides aminés ramifiée

Un peptide de la séquence ci-après :

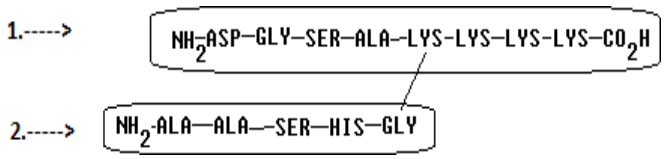


La liaison entre le résidu terminal de glycine de la séquence inférieure et la lysine de la séquence supérieure s'établit par l'intermédiaire d'une liaison amide entre la terminaison carboxyle de la glycine et la chaîne latérale du terminal amino de la lysine.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

La région non ramifiée ou linéaire d'une séquence contenant au moins quatre acides aminés définis de manière spécifique doit être intégrée dans un listage des séquences. Dans l'exemple ci-dessus, les régions linéaires du peptide ramifié qui comptent plus de quatre acides aminés sont les suivantes :



Le paragraphe 7.b) de la norme ST.26 prescrit l'intégration des séquences 1 et 2 dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Les séquences 1 et 2 doivent être représentés par des identificateurs de séquence distincts :

DGSAKKKK (SEQ ID NO : 32)

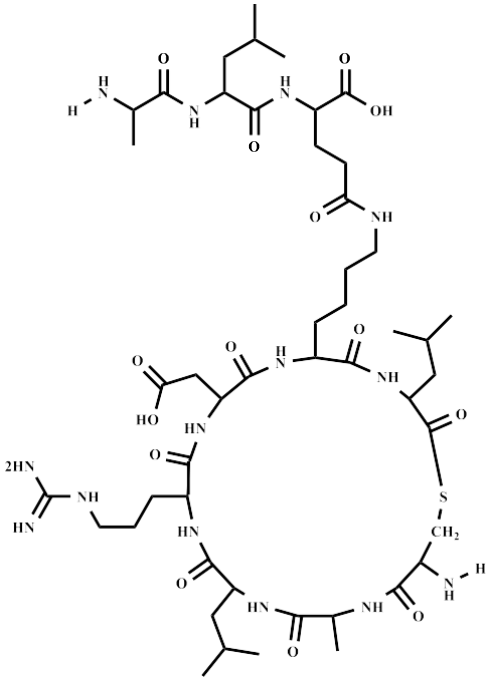
AASHG (SEQ ID NO : 33)

La séquence DGSAKKKK doit s'accompagner d'une annotation indiquant que la lysine en position numéro 5 est un acide aminé modifié; on utilisera la clé de caractérisation "SITE" ainsi que le qualificateur "note" pour indiquer que la chaîne latérale de la lysine est liée à une autre séquence par une liaison amide. La séquence AASHG devrait s'accompagner d'une annotation indiquant que la glycine en position numéro 5 est liée à une autre séquence en utilisant la clé de caractérisation "SITE" ainsi que le qualificateur "note".

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.b), 26, 29, 30 et 31

Exemple 7.b)-4 : Peptide cyclique contenant une séquence d'acides aminés ramifiée

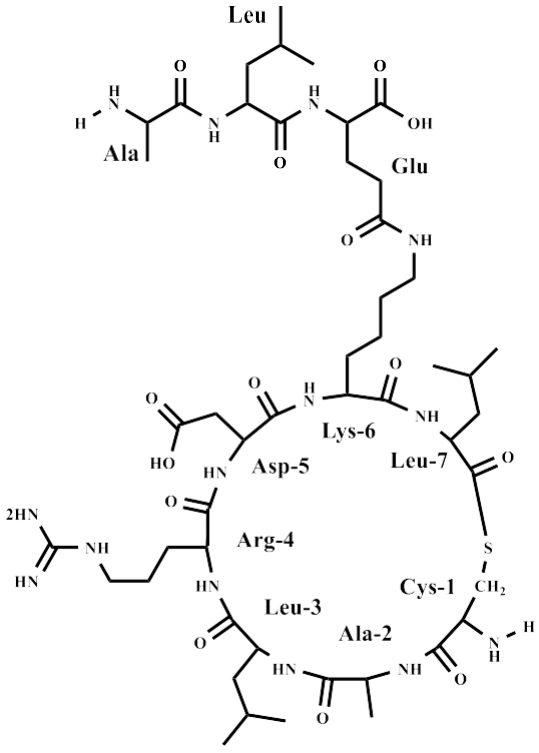
La structure suivante est divulguée dans une demande de brevet :



La cystéine et la leucine de la structure cyclique sont liées par la chaîne latérale de la Cys et la terminaison carboxy de la Leu.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

La structure ci-dessous est une séquence d'acides aminés cyclique et ramifiée contenant les acides aminés suivants :



Comme la chaîne latérale de la Cys et la terminaison carboxy de la Leu participent à la cyclisation, la terminaison N du peptide cyclique se trouve au Cys-1.

OUI –région cyclique du peptide

La norme ST.26 prévoit au paragraphe 7.b) que la région linéaire d'une séquence ramifiée contenant au moins quatre acides aminés définis de manière spécifique, et dont les acides aminés forment un seul squelette de peptides, doit figurer dans un listage des séquences. Dans l'exemple ci-dessus, la région cyclique du peptide ramifié compte plus de quatre acides aminés; elle doit donc être intégrée dans un listage des séquences.

NON – ramification tripeptide du peptide

Il n'est pas obligatoire d'intégrer la ramification du tripeptide Ala-Leu-Glu dans le listage des séquences.

Question 2 : La norme ST.26 autorise-t-elle l'intégration de la ou des séquences?

NON

Conformément au paragraphe 8, "un listage des séquences ne doit contenir aucune séquence comportant moins de quatre acides aminés définis de manière spécifique

La ramification du tripeptide ne contenant que trois acides aminés définis de manière spécifique, elle ne doit pas être intégrée dans le listage des séquences en tant que séquence distincte disposant de son propre numéro d'identification.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Bien que cet exemple illustre un peptide de configuration circulaire, l'anneau ne se compose pas uniquement de résidus d'acides aminés liés par des peptides, conformément au paragraphe 25. Comme la cyclisation de la séquence d'acides aminés s'effectue au moyen de la chaîne latérale de la cystéine (Cys) et de la terminaison carboxyle de la leucine (Leu), il faut attribuer le numéro de position 1 à la cystéine dans la région cyclique du peptide. La séquence doit donc être représentée comme suit :

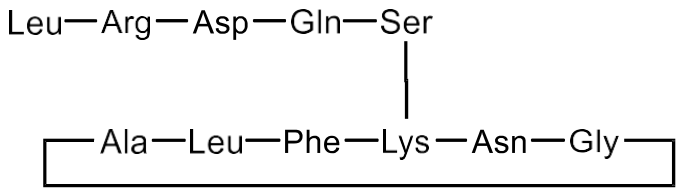
CALRDKL (SEQ ID NO :90).

Comme l'indique la figure ci-dessus, la séquence d'acides aminés est cyclisée par une conjugaison thioester entre la chaîne latérale de la cystéine et la terminaison carboxyle de la leucine. Il faut employer la clé de caractérisation "SITE" pour décrire la cystéine modifiée, qui forme la liaison intrachaîne avec la leucine. L'élément de l'emplacement de la caractéristique correspond aux numéros des résidus des acides aminés liés au format "x y, par exemple " 1.. 7". Le qualificateur obligatoire "note" doit indiquer la nature de la liaison, par exemple "cysteine leucine thioester (Cys-Leu)" pour préciser que le Cys-1 et le Leu-7 sont reliés par une liaison thioester. De plus, la lysine en position numéro 6 doit être annotée pour indiquer qu'elle a été modifiée; on emploie à cette fin la clé de caractérisation "SITE" ainsi que le qualificateur obligatoire "note", la valeur de ce dernier indiquant que le tripeptide ALE est relié à la chaîne latérale de la lysine.

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.b), 8, 25, 26, 29, 30, 31, 66.c) et 70

Exemple 7.b)-5 : Peptide cyclique contenant une séquence d'acides aminés ramifiée

Le peptide cyclique ramifié suivant est divulgué dans une demande de brevet :

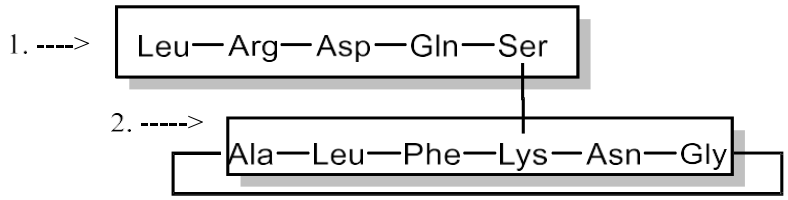


Le Ser et le Lys sont reliés par une liaison amide entre la terminaison carboxy de la sérine et l'amine dans la chaîne latérale du Lys.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Le paragraphe 7.b) prévoit que toute séquence contenant au moins quatre acides aminés définis de manière spécifique et pouvant être représentée comme une région linéaire dans une séquence ramifiée doit figurer dans le listage des séquences. Dans l'exemple ci-dessus, le peptide contient une région cyclique dans laquelle les acides aminés sont reliés par des liaisons peptidiques, ainsi qu'une région ramifiée qui est liée à une chaîne latérale du Lys dans la région cyclique. Les régions de ce peptide ramifié qui peuvent être représentées comme des régions linéaires et qui contiennent au moins quatre acides aminés définis de manière spécifique sont les suivantes :



La norme ST.26 prescrit d'intégrer les séquences 1 et 2 de ce peptide ramifié cyclique dans un listage des séquences, chacune de ces séquences devant disposer de son propre numéro d'identification.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence 1 doit être représentée ainsi :

LRDQS (SEQ. ID. NO :91)

La séquence 1 peut être annotée au moyen de la clé de caractérisation "SITE" et du qualificateur "note" pour indiquer que la sérine en position 5 est reliée à une autre séquence par une liaison amide entre le Ser et une chaîne latérale du Lys dans l'autre séquence.

La séquence 2 est un peptide cyclique. Selon le paragraphe 25, lorsqu'une séquence d'acides aminés a une configuration cyclique et n'a pas de terminaison amine ou carboxy, le requérant doit choisir le résidu d'acides aminés qu'il entend placer en position numéro 1. La séquence peut alors être représentée ainsi :

ALFKNG (SEQ. ID. NO :92)

Tout autre acide aminé de la séquence peut également être placé en position de résidu numéro 1. La séquence ALFKNG doit être décrite plus en détail au moyen de la clé de caractérisation "SITE" et du qualificateur "note" pour indiquer que la chaîne latérale du Lys en position de résidu numéro 4 est reliée par une liaison amide à une autre séquence. Cette liaison de la chaîne latérale modifie le Lys, et selon le paragraphe 30 de la norme ST.26, tout acide aminé modifié doit être accompagné d'une description supplémentaire dans le tableau de caractéristiques. De plus, une clé de caractérisation "REGION" et un qualificateur "note" devraient être prévus pour montrer que le peptide ALFKNG est circulaire.

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.b), 25, 26, 30 et 31

Paragraphe 11.a) – Séquence de nucléotides représentée par deux brins de codage – entièrement complémentaires

Exemple 11.a)-1 : Séquence de nucléotides représentée par deux brins de codage – mêmes longueurs

Une demande de brevet décrit la séquence d'ADN représentée par deux brins de codage ci-après :

3′-CCGGTTAACGCTA-5′

5′-GGCCAATTGCGAT-3′

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Chaque séquence nucléotidique énumérée comporte plus de 10 nucléotides définis de manière spécifique. Un brin de codage au moins doit être intégré dans le listage des séquences, car les deux brins de cette séquence nucléotidique représentée par deux brins de codage sont entièrement complémentaires l'un de l'autre.

Question 2 : La norme ST.26 autorise-t-elle l'intégration de la ou des séquences?

OUI

Si la séquence d'un seul brin de codage doit être intégrée dans le listage des séquences, les séquences des deux brins de codage peuvent y être intégrées, chacune disposant de son propre numéro d'identification de séquence.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence d'ADN représentée par deux brins de codage doit être représentée sous la forme d'une seule séquence ou de deux séquences distinctes. Chaque séquence intégrée dans le listage des séquences doit être représentée dans le sens 5'-3' et disposer de son propre numéro d'identification de séquence.

atcgcaattggcc (brin supérieur) (SEQ ID NO : 34)

et/ou

ggccaattgcgat (brin inférieur) (SEQ ID NO : 35)

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.a), **11.a)** et 13

Paragraphe 11.b) – Séquence de nucléotides représentée par deux brins de codage – non entièrement complémentaires

Exemple 11.b)-1 : Séquence de nucléotides représentée par deux brins de codage – différentes longueurs

Une demande de brevet contient le dessin et la légende ci-après :

```
5′-tagttcattgactaaggctccccattgactaaggcgactagcattgactaaggcaagc-3′
                    ||||||||||||||||
                    gggtaactgantccgc
```

Le promoteur du gène humain ABC1 (brin supérieur) lié par une sonde d'ANP (brin supérieur), où "n" dans la sonde d'ANP est une base d'ANP universelle choisie parmi le groupe comprenant le 5-nitroindole et le 3-nitroindole.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI – le promoteur de l'ABC1 (brin supérieur)

Contenant plus de 10 nucléotides énumérés et "définis de manière spécifique", le brin supérieur doit être intégré dans un listage des séquences.

OUI – la sonde ANP (brin inférieur)

Le brin inférieur doit également être intégré dans le listage des séquences en disposant de son propre numéro d'identification de séquence, car les deux brins ne sont pas entièrement complémentaires l'un de l'autre. Les résidus individuels qui constituent un ANP ou "acide nucléique peptidique" sont considérés comme des nucléotides conformément au paragraphe 3.g) de la norme ST.26. Il s'ensuit que le brin inférieur contient plus de 10 nucléotides énumérés et "définis de manière spécifique"; de ce fait, il doit être intégré dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Le brin supérieur doit être intégré dans un listage des séquences sous la forme suivante :

tagttcattgactaaggctccccattgactaaggcgactagcattgactaaggcaagc (SEQ ID NO : 36)

Le brin supérieur est un acide nucléique peptidique et, de ce fait, il ne possède pas d'extrémité 3' ni d'extrémité 5'. Conformément au paragraphe 11, il doit être intégré dans un listage des séquences "de gauche à droite de manière à reproduire le sens 5'-3'." Il doit donc être intégré dans un listage des séquences sous la forme suivante :

cgcctnagtcaatggg (SEQ ID NO : 37)

Le qualificateur du type "organism" de la clé de caractérisation "source" doit prendre la valeur "synthetic construct" et le qualificateur obligatoire du type "mol_type" la valeur "other DNA". Le brin inférieur doit être décrit dans un tableau de caractéristiques à l'aide de la clé de caractérisation "modified_base" et du qualificateur obligatoire "mod_base" avec l'abréviation "OTHER". Un qualificateur du type "note" doit être intégré avec le nom complet non abrégé des nucléotides modifiés, tels que les "N-(2-aminoethyl) glycine nucleosides".

Le résidu "n" doit s'accompagner d'une description supplémentaire dans un tableau de caractéristiques à l'aide de la clé de caractérisation "modified_base" et du qualificateur obligatoire "mod_base" avec l'abréviation "OTHER". Un qualificateur du type "note" doit être intégré avec le nom complet non abrégé du nucléotide modifié : "N-(2-aminoethyl) glycine 5-nitroindole ou N-(2-aminoethyl) glycine 3-nitroindole".

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.g), 7.a), **11.b)**, 17 et 18

Exemple 11.b)-2 : Séquence de nucléotides représentée par deux brins de codage – absence de segment d'appariement de bases

Une demande de brevet décrit la séquence d'ADN représentée par deux brins de codage ci-après :

```
3'-CCGGTTAGCTTATACGCTAGGGCTA-5'
   |||||||       ||||||||||||
5'-GGCCAATATGGCTTGCGATCCCGAT-3'
```

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Chaque brin de la séquence nucléotidique énumérée représentée par deux brins de codage contient plus de 10 nucléotides définis de manière spécifique. Les deux brins doivent être intégrés dans le listage des séquences, chacun disposant de son propre numéro d'identification de séquence, car les deux brins ne sont pas entièrement complémentaires l'un de l'autre.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence de chaque brin doit être représentée dans le sens 5'-3' et disposer de son propre numéro d'identification :

atcgggatcgcatattcgattggcc (brin supérieur) (SEQ ID NO : 38)

et

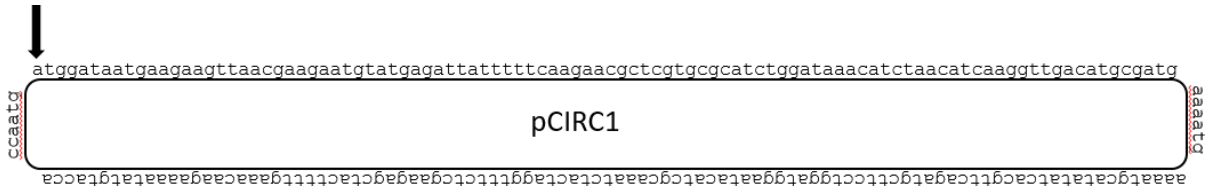
ggccaatatggcttgcgatcccgat (brin inférieur) (SEQ ID NO : 39)

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.a), **11.b)** et 13

Paragraphe 12 – Séquence nucléotidique circulaire

Example 12-1 : Séquence nucléotidique circulaire

Une demande de brevet contient la figure suivante, divulguant la séquence d'ADN du plasmide pCIRC1 :



Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

La séquence nucléotidique énumérée comporte plus de 10 nucléotides spécialement définis. Par conséquent, la séquence doit être intégrée dans un listage de séquences comme l'exige le paragraphe 7.a) de la norme ST.26.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Selon le paragraphe 12 de la norme ST.26, lorsque les séquences de nucléotides sont de configuration circulaire, le demandeur doit choisir le nucléotide en position de résidu numéro 1. Pour les besoins de cet exemple, le résidu "a" identifié par la flèche dans la figure sera utilisé comme position 1. Cependant, n'importe quel résidu peut être choisi comme position 1. Avec le résidu indiqué par la flèche comme position 1, la séquence doit être intégrée dans un listage de séquences comme suit :

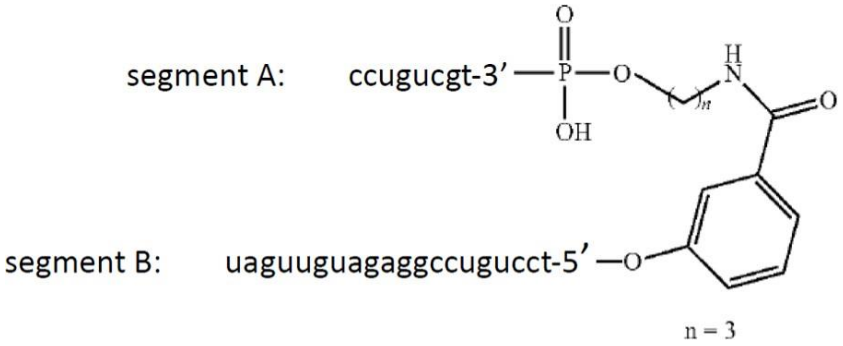
```
atggataatgaagaagttaacgaagaatgtatgagattatttttcaagaacgctcgtgcgcatctggataaacatctaaca
tcaaggttgacatgcgatgaaaatgaaaatgcatatatcacgttcagatgcttcctggatggaatacatcgcaaatctact
aggtttctcgaagagctacttttgaaacaagaaaatatgtaccaccaatg
```
(SEQ ID NO: 98)

La séquence doit être décrite plus en détail à l'aide de la clé de caractérisation "misc_feature" avec un emplacement de "212^1", qui indique que le dernier résidu de la séquence, la position 212, est lié au résidu 1. Le qualificateur "note" doit être inclus avec une valeur indiquant que la molécule est circulaire.

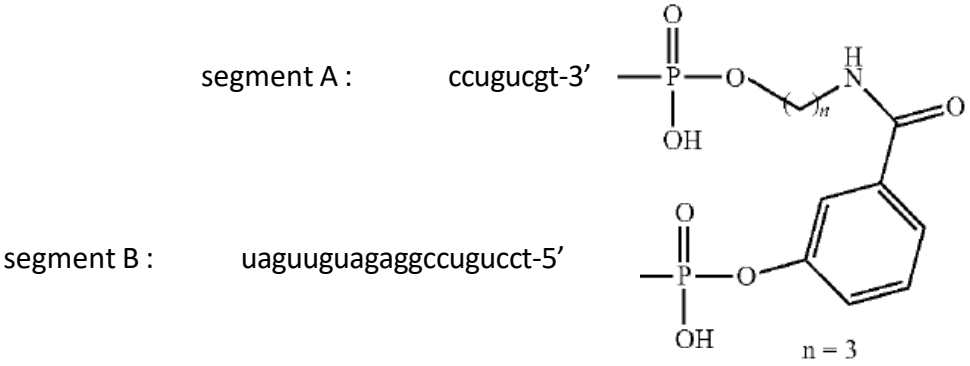
Paragraphe(s) pertinent(s) de la norme ST.26 : 7.a), **12** et annexe I, section 5, clé de caractérisation 5.15

Paragraphe 14 – Le symbole "t"désigne l'uracile dans de l'ARN

Exemple 14-1 : Le symbole "t" représente l'uracile dans de l'ARN



Une demande de brevet décrit le composé ci-après :



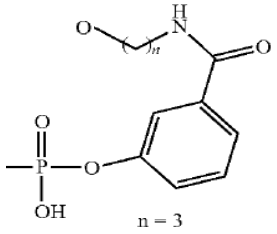
où le segment A et le segment B sont des séquences d'ARN.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

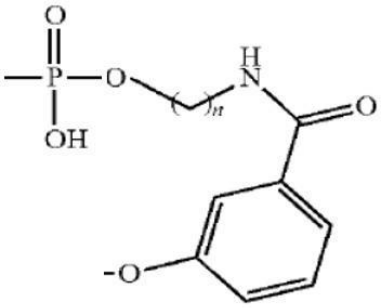
OUI – segment B

NON – segment A

La séquence énumérée contient deux segments de nucléotides définis de manière spécifique séparés par le "lieur" suivant :



n = 3



n = 3

Le lieur n'est pas un nucléotide au sens du paragraphe 3.g); de ce fait, chaque segment doit être considéré comme une séquence distincte. Le segment B contenant plus de 10 nucléotides définis de manière spécifique, le paragraphe 7.a) de la norme ST.26 en prescrit l'intégration dans un listage des séquences. Le segment A ne contenant que huit nucléotides définis de manière spécifique, la norme n'en prescrit pas l'intégration dans un listage des séquences.

Question 2 : La norme ST.26 autorise-t-elle l'intégration de la ou des séquences?

NON

Le segment A contenant moins de 10 nucléotides définis de manière spécifique et, selon le paragraphe 8 de la norme ST.26, il ne doit pas être intégré dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Le segment B est une molécule d'ARN; de ce fait, l'élément "INSDSeq_moltype" doit être "ARN." Le symbole "u" ne doit pas être utilisé pour représenter l'uracile dans une molécule d'ARN dans un listage des séquences. Selon le paragraphe 14, le symbole "t" désigne l'uracile dans de l'ARN. Le segment B doit donc être intégré dans le listage des séquences de la manière suivante :

tcctgtccggagatgttgat (SEQ ID NO : 40)

La thymine dans de l'ARN est considérée comme un nucléotide modifié, c'est-à-dire de l'uracile modifiée, et doit être représentée dans la séquence sous la forme "t" et s'accompagner d'une description supplémentaire dans un tableau de caractéristiques. En conséquence, la thymine à la position 1 doit s'accompagner d'une description supplémentaire à l'aide de la clé de caractérisation "modified_base", du qualificateur du type "mod_base" prenant la valeur "OTHER" et d'un qualificateur du type "note" prenant la valeur "thymine".

La thymine, c'est-à-dire l'uracile modifiée, à la position 1 devrait également s'accompagner d'une description supplémentaire dans un tableau de caractéristiques à l'aide de la clé de caractérisation "misc_feature" et d'un qualificateur du type "note" prenant, par exemple, la valeur " le phosphate 5' de la thymidine est fixé à une autre séquence de nucléotides par le lieur, qui est le3-hydroxybenzamido-N-propyl-3-phosphate. Chaque fois que possible, le qualificateur "note" peut directement prendre la valeur de l'autre séquence.

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.g), 7.a), 8, 13, **14**, 19 et 54

Paragraphe 27 – Il faut choisir le symbole ambigu le plus le plus restrictif

Exemple 27-1 : Formule topologique pour un acide aminé

$(GGGz)_2$

où z représente tout acide aminé.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

La séquence est divulguée sous la forme d'une formule. (GGGz)2 ne fait que représenter la séquence GGGzGGGz d'une manière abrégée. Par convention, on commence par amplifier une séquence et on détermine ensuite la définition de toute variable, à savoir "z".

La séquence utilise le symbole non conventionnel "z". La définition de "z" doit être déterminée à partir de l'explication de la séquence donnée dans la divulgation, qui définit ce symbole comme tout acide aminé (voir l'Introduction au présent document). L'exemple ne contient aucune contrainte pour "z", comme celle de devoir être identique à chaque occurrence.

Le peptide, dans cet exemple, comporte huit acides aminés énumérés, dont six sont des résidus de glycine définis de manière spécifique, et le reste des variables "z" qui devraient être représentées dans cette séquence au moyen du symbole conventionnel "X". Le paragraphe 7.b) de la norme ST.26 prescrit l'intégration de la séquence dans un listage des séquences sous la forme d'une séquence unique disposant d'un numéro d'identification de séquence unique.

On notera que la séquence reste visée par le paragraphe 7.b) en dépit du fait que les résidus énumérés et définis de manière spécifique ne sont pas contigus.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence utilise le symbole non conventionnel "z", qui, aux termes de la divulgation, représente tout acide aminé. Le symbole conventionnel utilisé pour représenter "tout acide aminé" est "X". Il s'ensuit que la séquence doit être représentée sous la forme de la séquence amplifiée unique suivante :

GGGXGGGX (SEQ ID NO : 41)

Conformément au paragraphe 27, "X" sera considéré comme l'équivalent de 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 dans le tableau de caractéristiques. Comme dans cet exemple, "X" représente "tout acide aminé", il doit être annoté au moyen de la clé de caractérisation "VARIANT" et d'un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

Chaque fois que possible, chaque "X" devrait être annoté individuellement. Cependant, une région de résidus "X" contigus ou un grand nombre de résidus "X" dispersés dans l'ensemble de la séquence peuvent être décrits conjointement par la clé de caractérisation "VARIANT" en employant la syntaxe "x..y" pour désigner le descripteur d'emplacement, où x et y sont les positions du premier et du dernier résidu "X", et par un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

D'autre part, l'exemple ne divulgue pas que "z" est le même acide aminé aux deux positions dans la séquence amplifiée. Toutefois, si "z" est divulgué comme le même acide aminé aux deux positions, il faudrait utiliser une clé de caractérisation "VARIANT" et un qualificateur du type "note" indiquant que "X" aux positions 4 et 8 peut représenter tous acides aminés dès l'instant qu'ils sont identiques aux deux positions.

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.c), 7.b) et 27

Exemple 27-2 : Formule topologique – moins de quatre acides aminés définis de manière spécifique

Un peptide de la formule (Gly-Gly-Gly-z)n

La divulgation indique également que z représente tout acide aminé et que

i) la variable n représente toute longueur; ou que

ii) la variable n est 2-100, de préférence 3

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

NON

L'examen des deux modes de réalisation divulgués i) et ii) du peptide énuméré de la formule montre que "n" peut représenter "toute longueur"; de ce fait, le mode de réalisation le plus englobant de "n" est indéterminé. Étant donné que "n" est indéterminé, le peptide de la formule ne peut pas être amplifié de manière à avoir une longueur précise, si bien qu'il faut prendre en compte la formule non amplifiée.

Le peptide énuméré dans la formule non amplifiée ("n" = 1) contient trois acides aminés définis de manière spécifique, dont chacun est Gly, et le symbole "z". Par convention, "Z" est le symbole représentant la "glutamine ou (l')acide glutamique"; toutefois, l'exemple définit "z" comme "tout acide aminé" (voir l'Introduction au présent document). En vertu de la norme ST.26, un acide aminé qui n'est pas défini de manière spécifique est représenté par "X". Il découle de cette analyse que le peptide énuméré, c'est-à-dire GGGX, ne contient pas quatre résidus d'acides aminés définis de manière spécifique. En conséquence, le paragraphe 7.b) de la norme ST.26 ne prescrit pas l'intégration, en dépit du fait que "n" est également défini comme représentant des valeurs numériques particulières dans certains modes de réalisation.

Question 2 : La norme ST.26 autorise-t-elle l'intégration de la ou des séquences?

OUI

L'exemple contient une valeur numérique particulière pour la variable "n", à savoir une limite inférieure de 2, une limite supérieure de 100 et une valeur exacte de 3. Toute séquence contenant au moins quatre acides aminés définis de manière spécifique peut être intégrée dans le listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Une séquence contenant 100 copies de GGGX est préférée (SEQ ID NO : 42). Il faudrait prévoir une annotation supplémentaire indiquant que l'on pourrait supprimer jusqu'à 98 copies de GGGX. L'intégration de modes de réalisation particuliers supplémentaires qui constituent une partie essentielle de l'invention est fortement conseillée.

Conformément au paragraphe 27, "X" sera considéré comme l'équivalent de 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 dans le tableau de caractéristiques. Étant donné que "X" dans la séquence SEQ ID NO: 42 représente "tout acide aminé", il doit être annoté à l'aide de la clé de caractérisation "VARIANT" et d'un qualificateur "note" prenant la valeur "X peut être tout acide aminé"..

Chaque fois que possible, chaque "X" devrait être annoté individuellement. Cependant, une région de résidus "X" contigus ou un grand nombre de résidus "X" dispersés dans l'ensemble de la séquence peuvent être décrits conjointement par la clé de caractérisation "VARIANT" en employant la syntaxe "x.y" pour désigner le descripteur d'emplacement, où x et y sont les positions du premier et du dernier résidu "X", et par un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

N.B. : La représentation préférée de la séquence indiquée ci-dessus sert à fournir un listage des séquences à la date du dépôt d'une demande de brevet. La même représentation pourra ne pas être applicable à un listage des séquences fourni après cette date, car il faut tenir compte de la question de savoir si l'information fournie pourrait être prise en considération par un office de la propriété intellectuelle pour ajouter des éléments à la divulgation originale.

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.c), 7.b), 26 et **27**

Exemple 27-3 : Formule topologique – au moins quatre acides aminés définis d'une manière spécifique

Un peptide de la formule (Gly-Gly-Gly-z)n
où z est tout acide aminé et la variable n est compris entre 2 et 100, et est de préférence 3.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Le peptide énuméré dans la formule contient trois acides aminés définis de manière spécifique, dont chacun est Gly, et le symbole "z". Par convention, "Z" est le symbole représentant la "glutamine ou (l')acide glutamique"; toutefois, la description, dans cet exemple, définit "z" comme "tout acide aminé" (voir l'Introduction au présent document). En vertu de la norme ST.26, un acide aminé qui n'est pas défini de manière spécifique est représenté par "X". Il découle de cette analyse que le peptide répété énuméré ne contient pas quatre acides aminés définis de manière spécifique. Toutefois, la description indique une valeur numérique particulière pour la variable "n", c'est-à-dire une limite inférieure de 2 et une limite supérieure de 100. En conséquence, l'exemple divulgue un peptide contenant aux moins six acides aminés définis de manière spécifique dans la séquence GGGzGGGz, qui doit, conformément à la norme ST.26, être intégrée dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Étant donné que "z" représente tout acide aminé, le symbole conventionnel utilisé pour représenter le quatrième et le huitième acide aminé est "X."

La norme ST.26 prescrit l'intégration dans un listage des séquences de la seule séquence qui a été énumérée par ses résidus. De ce fait, au moins une séquence contenant 2, 3 ou 100 copies de GGGX doit être intégrée dans le listage des séquences; toutefois, la séquence la plus englobante qui contient 100 copies de GGGX est préférée (SEQ ID NO : 42) (voir l'Introduction au présent document). Dans ce dernier cas, une annotation supplémentaire pourrait indiquer que 98 copies de GGGX au maximum pourraient être supprimées. Il est fortement conseillé d'intégrer deux séquences supplémentaires contenant 2 et 3 copies de GGGX, respectivement (SEQ ID NO : 44-45).

Conformément au paragraphe 27, "X" sera considéré comme l'équivalent de 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 dans le tableau de caractéristiques. Comme "X", dans cet exemple, représente "tout acide aminé", il doit être annoté au moyen de la clé de caractérisation "VARIANT" et d'un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

Chaque fois que possible, chaque "X" devrait être annoté individuellement. Cependant, une région de résidus "X" contigus ou un grand nombre de résidus "X" dispersés dans l'ensemble de la séquence peuvent être décrits conjointement par la clé de caractérisation "VARIANT" en employant la syntaxe "x..y" pour désigner le descripteur d'emplacement, où x et y sont les positions du premier et du dernier résidu "X", et par un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

D'autre part, l'exemple ne divulgue pas que la variable "z" est la même dans chacune des deux occurrences dans la séquence amplifiée. Toutefois, si "z" est divulgué comme étant le même acide aminé dans tous les emplacements, une clé de caractérisation VARIANT et un qualificateur du type note devraient indiquer que "X" à toutes les positions peut représenter tous acides aminés, dès l'instant qu'ils sont les mêmes à tous les emplacements.

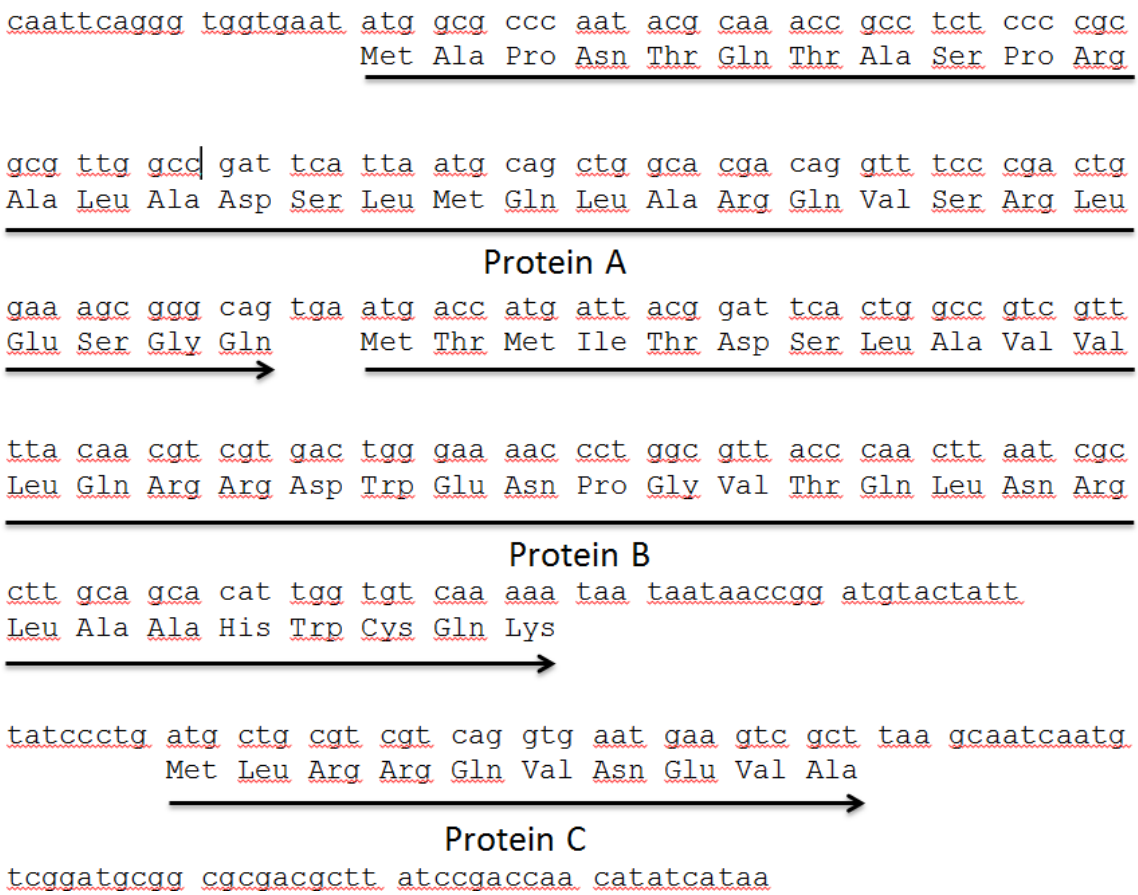
N.B. La représentation préférée de la séquence indiquée ci-dessus sert à fournir un listage des séquences à la date du dépôt d'une demande de brevet. La même représentation pourra ne pas être applicable à un listage des séquences fourni après cette date, car il faut tenir compte de la question de savoir si l'information fournie pourrait être prise en considération par un office de la propriété intellectuelle pour ajouter des éléments à la divulgation originale.

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.c), 7.b), 26, et **27**

Paragraphe 28 – Séquences d'acides aminés séparées par des symboles internes de fin

Exemple 28-1 : Codage de la séquence de nucléotides et séquence d'acides aminés codée

Une demande de brevet décrit les séquences ci-après :



Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

La demande décrit une séquence nucléotidique contenant des codons de terminaison et codant trois séquences d'acides aminés distinctes.

Contenant plus de 10 nucléotides définis de manière spécifique, la séquence nucléotidique énumérée doit être intégrée dans un listage des séquences comme une séquence unique.

En ce qui concerne les séquences d'acides aminés codées, le paragraphe 28 prescrit l'intégration comme séquences distinctes des séquences d'acides aminés séparées par un symbole interne de fin tel qu'un espace blanc. Étant donné que chaque "Protéine A", "Protéine B" et "Protéine C" contient au moins quatre acides aminés définis de manière spécifique, elles doivent chacune, conformément au paragraphe 7.b) de la norme ST.26, être intégrées dans un listage des séquences et disposer de leur propre numéro d'identification de séquence.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence nucléotidique doit être intégrée dans un listage des séquences sous la forme suivante :

caattcagggtggtgaatatggcgcccaatacgcaaaccgcctctccccgcgcgttggccgattcattaatgcagctggccaggcaggtgagcaggctgg
aaagcgggcagtgaatgaccatgattacggattcactggccgtcgttttacaacgtcgtgactgggaaaaccctggcgttacccaacttaatcgccttgcag
cacattggtgtcaaaaataataataaccggatgtactatttatccctgatgctgcgtcgtcaggtgaatgaagtcgcttaagcaatcaatgtcggatgcggcg
cgacgcttatccgaccaacatatcataa (SEQ ID NO : 46)

La séquence nucléotidique devrait s'accompagner d'une description supplémentaire utilisant une clé de caractérisation "CDS" pour chacune des trois protéines, et l'élément INSDFeature_location doit indiquer l'emplacement de chaque séquence de codage, y compris le codon d'arrêt. De plus, il faudrait employer, en association avec chaque clé de caractérisation "CDS", le qualificateur "translation" qui prenne comme valeur la séquence d'acides aminés de la protéine. La demande ne divulgue pas le tableau du code génétique qui est appliqué à la traduction (voir annexe 1, section 9, tableau 7). Si le tableau de codes normalisés est appliqué, le qualificateur "transl_table" n'est pas nécessaire; toutefois, si un tableau du code génétique différent est appliqué, il faut indiquer la valeur appropriée figurant dans le tableau 7 pour le qualificateur "transl_table". Enfin, il faut intégrer le qualificateur "protein _id" dont la valeur indiquera le numéro d'identification de séquence de chacune des séquences d'acides aminés traduites.

Les séquences d'acides aminés doivent être intégrées comme séquences distinctes disposant chacune de leur propre numéro d'identification de séquence :

MAPNTQTASPRALADSLMQLARQVSRLESGQ (SEQ ID NO : 47)

MTMITDSLAVVLQRRDWENPGVTQLNRLAAHWCQK (SEQ ID NO : 48)

MLRRQVNEVA (SEQ ID NO : 49)

NOTE : Voir "Exemple 90-1 Séquence d'acides aminés codée selon une séquence de codage avec introns" pour un exemple de séquence d'acides aminés traduite représentée en tant que séquence unique.

P**aragraphe(s) pertinent(s) de la norme ST.26 :** 7, 26, **28**, 57, 89 à 92

Paragraphe 29 – Représentation d'un acide aminé modifié "other"

Exemple 29-1 : Symbole ambigu le plus restrictif pour un acide aminé "other"

Une demande de brevet décrit la séquence ci-après :

Ala-Hse-X_1-X_2-X_3-X_4-Tyr-Leu-Gly-Ser

où, X1= Ala ou Gly,

X2= Ala ou Gly,

X3= Ala ou Gly,

X4= Ala ou Gly, et

Hse = homosérine

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Le peptide énuméré contient cinq acides aminés définis de manière spécifique. Le symbole "X" est employé par convention pour représenter deux acides aminés alternatifs (voir l'Introduction au présent document).

Étant donné qu'il y a cinq acides aminés définis de manière spécifique, à savoir, Ala, Tyr, Leu, Gly et Ser, le paragraphe 7.b) de la norme ST.26 prescrit l'intégration de la séquence dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

En vertu du paragraphe 29, tout acide aminé "other" doit être représenté par le symbole "X". Dans l'exemple, la séquence contient l'acide aminé Hse à la position 2, qui ne figure pas dans l'annexe I, section 3, tableau 3. Il s'ensuit que Hse est un acide aminé "other" et doit être représenté par le symbole "X".

X_1-X_4 sont des positions variantes, chacune d'elles pouvant correspondre à A ou à G. Le symbole ambigu le plus restrictif représentant les acides aminés alternatifs A ou G est "X". La séquence peut donc être représentée sous la forme suivante :

AXXXXXYLGS (SEQ ID NO : 50)

Comme indiqué dans l'introduction au présent document, l'intégration de toutes séquences particulières de première importance pour la divulgation ou les revendications de l'invention est fortement conseillée.

Étant donné que l'acide aminé Hse ne figure pas dans l'annexe I, section 4, tableau 4, une clé de caractérisation "SITE" et un qualificateur "note" doivent être employés et doivent indiquer le nom complet non abrégé de l'homosérine, conformément au paragraphe 30 de la norme ST.26.

Selon le paragraphe 27, puisque X_1-X_4 représentent un choix entre deux acides aminés seulement, une description supplémentaire s'impose. Le paragraphe 96 indique qu'il conviendrait d'utiliser la clé de caractérisation "VARIANT" avec le qualificateur "note", qui prendrait la valeur "A ou G". Selon le paragraphe 34 de la norme ST.26, comme elles sont adjacentes et décrites de manière identique, ces positions peuvent faire l'objet d'une description commune à l'aide de la syntaxe "3...6" en tant que descripteur d'emplacement de l'élément INSDFeature_location.

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.a), 7.b), 25 à 27, **29**, 30 34, 66, 70, 71 et 96-97

Exemple 29-2 : Utilisation de l'acide aminé non modifié correspondant

Une demande de brevet décrit la séquence suivante :

Ala-Hyl-Tyr-Leu-Gly-Ser-Nle-Val-Ser-5ALA

Où Hyl = hydroxylysine (modification post-traductionnelle de la lysine), Nle = Norleucine, et 5ALA = acide δ-aminolévulinique

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Le peptide énuméré contient plus de quatre acides aminés définis de manière spécifique; par conséquent, la séquence doit être intégrée dans un listage de séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

L'hydroxylysine à la position 2, la norleucine à la position 7 et l'acide δ-aminolévulinique à la position 10 sont tous des "acides aminés modifiés". En premier lieu, nous devons examiner chaque acide aminé modifié et déterminer s'il devrait être représenté par l'acide aminé non modifié correspondant ou par la variable "X" dans la séquence. Le paragraphe 29 indique qu'un acide aminé modifié "devrait être représenté dans la séquence comme l'acide aminé non modifié correspondant chaque fois que possible".

Il est laissé à la discrétion du déposant de décider si un acide aminé modifié sera représenté par un résidu non modifié correspondant ou par les variables "X". Les conseils suivants devraient toutefois être pris en considération : si un acide aminé est modifié par l'ajout d'une fraction, telle que la méthylation ou l'acétylation, et que la structure de l'acide aminé non modifié correspondant demeure généralement inchangée, la représentation par un acide aminé non modifié est alors recommandée. Si l'acide aminé modifié est structurellement très différent de l'acide aminé non modifié correspondant, alors la représentation par un "X" est recommandée.

La structure de l'hydroxylysine est presque identique à la lysine, à l'exception du fait que le troisième carbone dans le groupe-R est modifié au moyen d'un groupe hydroxyle. Étant donné que la structure de base du résidu de lysine non modifié correspondant est intacte, l'hydroxylysine devrait être représentée dans la séquence par la lysine ("K"), et non par "X".

La norleucine est un isomère de la leucine. Le groupe R de la leucine est une chaîne de 4 carbones, ramifiée au niveau du deuxième carbone. La norleucine comprend également un groupe R de 4 carbones mais qui est linéaire et n'est pas ramifié. Par conséquent, la norleucine n'est pas simplement le résultat d'une modification ajoutée à la leucine, mais une structure complètement différente (bien qu'apparentée). Il est donc recommandé de représenter la norleucine par un "X" dans un listage des séquences.

L'acide δ-aminolévulinique n'est structurellement similaire à aucun des acides aminés répertoriés dans le tableau 3 de l'annexe I. Il est donc recommandé de représenter l'acide δ-aminolévulinique par un "X" dans un listage des séquences.

Par conséquent, la séquence devrait être intégrée dans un listage des séquences sous la forme :

AKYLGSXVSX (SEQ ID NO.51)

Le paragraphe 30 prescrit que chaque acide aminé modifié doit être accompagné d'une description supplémentaire.

L'hydroxylysine est une modification post-traductionnelle de la lysine. Par conséquent, elle doit être décrite à l'aide de la clé de caractérisation "MOD_RES" et du qualificateur "note" qui précise la modification. Il convient de noter que l'"hydroxylysine" est répertoriée dans le tableau 4 de l'annexe I, section 4, intitulé "Liste des acides aminés modifiés". La valeur du qualificateur "note" peut donc contenir l'abréviation "Hyl" au lieu du nom non abrégé "hydroxylysine".

La norleucine n'est pas un résidu modifié après traduction et doit par conséquent être décrite à l'aide de la clé de caractérisation "SITE" et du qualificateur "note" qui précise la modification. Il convient de noter que la "norleucine" est également répertoriée dans le tableau 4 de la section 4 de l'annexe 1. Par conséquent, la valeur du qualificateur "note" peut contenir l'abréviation "Nle" au lieu du nom non abrégé "norleucine".

L'acide δ-aminolévulinique n'est pas un résidu modifié après traduction et doit donc être décrit à l'aide de la clé de caractérisation "SITE" et du qualificateur "note" qui précise la modification. L'acide δ-aminolévulinique n'est pas répertorié dans le tableau 4 de la section 4 de l'annexe 1 et la valeur du qualificateur "note" doit donc contenir le nom complet non abrégé du résidu modifié "acide δ-aminolévulinique".

Paragraphes pertinents de la norme ST.26 : 3.a), 3.e), 7.b), 29 et 30

Paragraphe 30 – Annotation d'un acide aminé modifié

Exemple 30-1 – Clé de caractérisation "CARBOHYD"

Une demande de brevet décrit un polypeptide contenant un acide aminé modifié de manière spécifique, dont une chaîne latérale a fait l'objet d'une glycosylation et qui se caractérise par le fait que la Cys correspondant aux positions 4 et 15 du polypeptide forme une liaison disulfure, conformément à la séquence ci-après :

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

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Le peptide énuméré contient 17 acides aminés définis de manière spécifique. Il y a 16 acides aminés naturels, dont le neuvième (asparagine) a fait l'objet d'une glycosylation. La séquence doit donc être intégrée dans un listage des séquences comme le prescrit le paragraphe 7.b) de la norme ST.26.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Aux termes du paragraphe 29 de la norme ST.26, un acide aminé modifié devrait être représenté dans la séquence comme l'acide aminé non modifié correspondant chaque fois que possible.

La séquence doit donc être intégrée dans un listage des séquences sous la forme suivante :

LEYCLKRWNETISHCAW (SEQ ID NO : 52)

L'acide aminé modifié doit s'accompagner d'une description supplémentaire. La clé de caractérisation "CARBOHYD" et le qualificateur (obligatoire) du type "note" doivent être employés pour indiquer l'occurrence de la fixation d'une chaîne de sucre (asialyloligosaccharide) sur l'asparagine à la position 9. Le qualificateur du type "note" décrit le type de liaison, par exemple une N-liaison. Le descripteur d'emplacement dans l'élément de l'emplacement de la caractéristique est le numéro de position de résidu de l'asparagine modifiée.

En outre, il y a une liaison disulfure entre les deux résidus de Cys. De ce fait, la clé de caractérisation "DISULFID" devrait donc être utilisée pour décrire une liaison intrachaîne. L'élément de l'emplacement de la caractéristique correspond aux numéros de position des résidus de Cys liés au format "x y", c'est-à-dire "4 15". Le qualificateur obligatoire du type note devrait décrire la liaison intrachaîne disulfure.

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.a), 7.b), 26, 29, 30, 66.c), 70 et annexe I, section 7, clé de caractérisation 7.4

Exemple 30-2 – Acides aminés modifiés après traduction

Une demande de brevet décrit le polypeptide suivant :

Leu-Glu-Tyr-Cys-Leu-Lys-Arg-Trp-Xaa-Glu-Thr-Ile-Ser-His

dans lequel l'Arg en position 7 est modifié en citrulline après traduction.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Le peptide énuméré contient 13 acides aminés spécialement définis. Par conséquent, la séquence doit être intégrée dans un listage des séquences comme l'exige le paragraphe 7.b) de la norme ST.26.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Selon le paragraphe 29 de la norme ST.26, un acide aminé modifié doit être représenté dans la séquence comme l'acide aminé non modifié correspondant chaque fois que cela est possible.

Par conséquent, la séquence doit être intégrée dans un listage de séquences comme suit :

LEYCLKRWXETISHCAW (SEQ ID NO: xx)

où le symbole "R" est utilisé pour représenter l'arginine en position 7.

Une description supplémentaire indiquant que l'arginine en position 7 peut être modifiée en citrulline est requise. La modification de l'arginine en citrulline est une modification après traduction. Par conséquent, la clé de caractérisation "MOD_RES" doit être utilisée avec le qualificateur obligatoire "note" pour indiquer que l'arginine peut être modifiée pour devenir de la citrulline. Dans l'emplacement d'une caractéristique, le descripteur d'emplacement est le numéro de position du résidu de l'arginine modifiée.

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.a), 7.b), **30** et annexe I, section 7, clé de caractérisation 7.18

Paragraphe 36 – Séquences contenant des régions comportant un nombre exact de résidus contigus "n" ou "X"

Exemple 36-1 : Séquence dont une région contient un nombre connu de résidus "X" représentée comme une séquence unique

LL-100-KYMR

où "-100-"entre les acides aminés leucine et lysine correspond à une région de 100 acides aminés dans la séquence.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Le paragraphe 36 de la norme ST.26 prescrit l'intégration d'une séquence qui contient au moins quatre acides aminés définis de manière spécifique et séparés par une ou plusieurs régions contenant un nombre déterminé de résidus "X".

La séquence divulguée utilise un symbole non conventionnel, à savoir "-100-." La définition de "-100-" doit être tirée de l'explication de la séquence donnée dans la divulgation, qui définit ce symbole comme 100 acides aminés entre la leucine et la lysine (voir Introduction au présent document). De ce fait, "-100-" est une région déterminée de résidus "X". Étant donné que six des 106 acides aminés de la séquence sont définis de manière spécifique, le paragraphe 7.b) de la norme ST.26 prescrit l'intégration de la séquence dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Le symbole non conventionnel "-100-" est représenté comme 100 résidus "X" (car tout symbole utilisé pour représenter un acide aminé est l'équivalent d'un seul résidu). Il s'ensuit qu'une séquence unique de 106 acides aminés de longueur, contenant 100 résidus "X" entre LL et KYMR, doit être intégrée dans un listage des séquences (SEQ ID NO : 53).

Cette séquence contient 100 variables "X" entre LL et KYMR. Dans la norme ST.26, le symbole "X" non accompagné d'une description supplémentaire est par défaut 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" (paragraphe 27). Si ces 100 variables "X" sont définies comme prenant une autre valeur que cette valeur par défaut, alors une description supplémentaire appropriée pour chaque variable "X" doit être indiquée.

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.b), 26, 27 et **36**

Exemple 36-2 : Séquence dont plusieurs régions contiennent un nombre ou une série connu de résidus "X" représentée comme une séquence unique

Lys-z_2-Lys-z_m-Lys-z_3-Lys-z_n-Lys-z_2-Lys

où z représente tout acide aminé, m=20, n=19-20, z_2 signifie que les paires de lysine sont séparées par deux acides aminés quels qu'ils soient et z_3 signifie que les paires de lysine sont séparées par trois acides aminés quels qu'ils soient.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

La séquence divulguée utilise un symbole non conventionnel, à savoir "z." Il faut donc consulter la divulgation pour déterminer la définition; "z" est défini comme tout acide aminé (voir l'introduction au présent document). Le symbole conventionnel utilisé pour représenter tout acide aminé est "X". Compte tenu de la présence de variables "X", le peptide contient six résidus de lysine qui sont énumérés et définis de manière spécifique, à intégrer dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence utilise un symbole non conventionnel "z", dont la définition doit être tirée de la divulgation. Étant donné que "z" est défini comme tout acide aminé, le symbole conventionnel est "X."

Il est préférable d'utiliser le moyen de représentation le plus général, qui est le suivant (voir l'introduction au présent document) :

KXXKXXXXXXXXXXXXXXXXXXXXKXXXKXXXXXXXXXXXXXXXXXXXXXXKXXK (SEQ ID NO : 54)

où z_n est égal à 20 "X's", une description supplémentaire indiquant que la variable "X" correspondant à la position 30 peut être supprimée.

En lieu et place ou en sus de ce qui précède, la séquence peut être représentée sous la forme suivante :

KXXKXXXXXXXXXXXXXXXXXXXXKXXXKXXXXXXXXXXXXXXXXXXXXXXKXXK (SEQ ID NO : 55)

où z_n est égal à 19 "X's", une description supplémentaire indiquant qu'une variable "X" peut être ajoutée entre les numéros de position 29 et 30.

Conformément au paragraphe 27, "X" sera considéré comme l'équivalent de 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 dans le tableau de caractéristiques. Étant donné que dans les séquences SEQ ID NO : 54 et SEQ ID NO : 55 , "X" représente "tout acide aminé", il doit être annoté au moyen de la clé de caractérisation "VARIANT" et d'un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

Chaque fois que possible, chaque "X" devrait être annoté individuellement. Cependant, une région de résidus "X" contigus ou un grand nombre de résidus "X" dispersés dans l'ensemble de la séquence peuvent être décrits conjointement par la clé de caractérisation "VARIANT" en employant la syntaxe "x...y" pour désigner le descripteur d'emplacement, où x et y sont les positions du premier et du dernier résidu "X", et par un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

Paragraphe(s) pertinent(s) de la norme ST.26 : 26, 27 et **36**

Exemple 36-3 : Séquence dont plusieurs régions contiennent un nombre ou une série connus de résidus "X" et qui est représentée comme une séquence unique

K-z2-K-zm-K-z3-K-zn-K-z2-K

où z est tout acide aminé, m=15-25, de préférence 20-22, n=15-25, de préférence 19-20, z2 signifie que les paires de lysine sont séparées par deux acides aminés quels qu'ils soient et z3 signifie que les paires de lysine sont séparées par trois acides aminés quels qu'ils soient.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

La séquence de l'exemple utilise un symbole non conventionnel, à savoir "z". De ce fait, on consulte la divulgation pour déterminer la définition de "z" (voir l'Introduction au présent document). La divulgation définit ce symbole comme tout acide aminé. Le symbole conventionnel utilisé pour représenter un résidu défini comme "tout acide aminé" est "X." Compte tenu de la présence de résidus "X", le peptide contient six résidus de lysine qui sont énumérés et définis de manière spécifique, à intégrer dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence utilise un symbole non conventionnel "z", dont la définition doit être tirée de la divulgation. Étant donné que "z" est défini comme tout acide aminé, le symbole conventionnel est "X". Le moyen de représentation à préférer et le plus englobant est le suivant :

KXXKXXXXXXXXXXXXXXXXXXXXXXXXXKXXXKXXXXXXXXXXXXXXXXXXXXXXXXXKXXK (SEQ ID NO : 56)

(où m=25 et n=25), une description supplémentaire indiquant que l'on peut supprimer jusqu'à 10 résidus "X" dans chacune des régions "zm" ou "zn".

Comme indiqué dans l'introduction au présent document, l'intégration de toutes séquences particulières de première importance pour la divulgation ou les revendications de l'invention est fortement conseillée.

La séquence pourrait également être représentée sous la forme suivante :

KXXKXXXXXXXXXXXXXXXKXXXKXXXXXXXXXXXXXXXKXXK (SEQ ID NO : 57)

(où m=15 et n=15), une description supplémentaire indiquant que l'on peut ajouter jusqu'à 10 résidus "X" dans chacune des régions "zm" ou "zn".

Toutes autres variantes possibles pourraient également être intégrées.

Conformément au paragraphe 27, "X" sera considéré comme l'équivalent de 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 dans le tableau de caractéristiques. Étant donné que, dans les séquences SEQ ID NO : 56 et SEQ ID NO : 57, "X" représente "tout acide aminé", il doit être annoté au moyen de de la clé de caractérisation "VARIANT" et d'un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

Chaque fois que possible, chaque "X" devrait être annoté individuellement. Cependant, une région de résidus "X" contigus ou un grand nombre de résidus "X" dispersés dans l'ensemble de la séquence peuvent être décrits conjointement par la clé de caractérisation "VARIANT" en employant la syntaxe "x..y" pour désigner le descripteur d'emplacement, où x et y sont les positions du premier et du dernier résidu "X", et par un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

N.B. La représentation préférée de la séquence indiquée ci-dessus sert à fournir un listage des séquences à la date du dépôt d'une demande de brevet. La même représentation pourra ne pas être applicable à un listage des séquences fourni après cette date, car il faut tenir compte de la question de savoir si l'information fournie pourrait être prise en considération par un office de la propriété intellectuelle pour ajouter des éléments à la divulgation originale.

Paragraphe(s) pertinent(s) de la norme ST.26 : 27 et **36**

Paragraphe 37 – Séquences dont des régions contiennent un nombre inconnu de résidus contigus "n" ou "X"

Exemple 37-1 : Une séquence dont des régions contiennent un nombre inconnu de résidus "X" ne doit pas être représentée comme une séquence unique

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

où le symbole -----est une brèche non définie dans la séquence, Xaa représente tout acide aminé, et les résidus de glycine et de Xaa sont reliés entre eux par des liaisons peptidiques.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

NON

Le paragraphe 37 de la norme ST.26 interdit l'intégration d'une séquence qui contient une brèche non définie; l'intégration de l'ensemble de la séquence n'est donc pas prescrite.

En revanche, le paragraphe 37 de la norme ST.26 prescrit l'intégration de toute région de séquence adjacente à une brèche non définie qui contient au moins quatre acides aminés définis de manière spécifique. Dans l'exemple ci-dessus, l'intégration de toute région de séquence adjacente à une brèche non définie n'est pas prescrite, car chaque région ne contient que deux acides aminés définis de manière spécifique.

Question 2 : La norme ST.26 autorise-t-elle l'intégration de la ou des séquences?

NON – pas l'ensemble de la séquence

NON – aucune région de la séquence

Le paragraphe 37 de la norme ST.26 n'autorise pas l'intégration de l'ensemble de la séquence.

Le paragraphe 8 de la norme ST.26 n'autorise pas l'intégration de l'une ou de l'autre des régions adjacentes à la brèche non définie, car chaque région ne contient que deux acides aminés définis de manière spécifique.

Paragraphe(s) pertinent(s) de la norme ST 26 : 7.b), 8, 26 et **37**

Exemple 37-2 : Une séquence dont des régions contiennent un nombre inconnu de résidus "X" ne doit pas être représentée comme une séquence unique

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

où le symbole ----- est une brèche non définie dans la séquence, Xaa représente tout acide aminé, et les résidus de glycine et de Xaa sont reliés entre eux par des liaisons peptidiques.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

NON – pas l'ensemble de la séquence

OUI – une région de la séquence

Le paragraphe 37 de la norme ST.26 interdit l'intégration d'une séquence qui contient une brèche non définie, mais prescrit l'intégration de toute région de séquence adjacente à une brèche non définie qui contient au moins quatre acides aminés définis de manière spécifique.

Dans l'exemple ci-dessus, la norme ST.26 ne prescrit pas (et interdit) l'intégration de l'ensemble de la séquence, qui contient une brèche non définie, et de la région Gly-Gly adjacente à ladite brèche, qui ne contient que deux acides aminés définis de manière spécifique. Toutefois, la norme ST.26 prescrit l'intégration de la région Gly-Gly-Ala-Gly- Xaa-Xaa adjacente à la brèche non définie, car elle contient au moins quatre acides aminés définis de manière spécifique.

Question 2 : La norme ST.26 autorise-t-elle l'intégration de la ou des séquences?

NON – pas de l'ensemble de la séquence ni de la région Gly-Gly

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La région de la séquence adjacente à la brèche non définie qui contient quatre acides aminés définis de manière spécifique doit être représentée sous la forme suivante :

GGAGXX (SEQ ID NO : 58)

La séquence devrait faire l'objet d'une annotation indiquant que la séquence représentée fait partie d'une séquence plus longue qui contient une brèche non définie en utilisant la clé de caractérisation "SITE", l'emplacement de caractéristique "1" et le qualificateur du type "note" qui prendrait par exemple la valeur "Ce résidu est relié par une liaison N-terminale à un peptide comportant un Gly-Gly N-terminal et une brèche de longueur non définie."

Conformément au paragraphe 27, "X" sera considéré comme l'équivalent de 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 dans le tableau de caractéristiques. Étant donné que "X" dans la séquence SEQ ID NO : 58 représente "tout acide aminé", il doit être annoté au moyen de la clé de caractérisation "VARIANT" et d'un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

Chaque fois que possible, chaque "X" devrait être annoté individuellement. Cependant, une région de résidus "X" contigus ou un grand nombre de résidus "X" dispersés dans l'ensemble de la séquence peuvent être décrits conjointement par la clé de caractérisation "VARIANT" en employant la syntaxe "x..y" pour désigner le descripteur d'emplacement, où x et y sont les positions du premier et du dernier résidu "X", et par un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

Paragraphe(s) pertinent(s) de la norme ST 26 : 7.b), 8, 26, 27 et **37**

Paragraphe 55 – Séquence de nucléotides contenant à la fois des segments d'ADN et d'ARN

Exemple 55-1 : Molécule combinée d'ADN et d'ARN

La séquence d'oligonucléotides suivante est divulguée dans une demande de brevet :

AGACCTTcggagucuccuguugaacagauagucaaaguagauC

où les majuscules représentent des résidus d'ADN et les minuscules des résidus d'ARN.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

La séquence divulguée contient plus de 10 nucléotides énumérés et définis de manière spécifique; elle doit donc être intégrée dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence nucléotidique doit être intégrée dans un listage des séquences sous la forme suivante :

agaccttcggagtctcctgttgaacagatagtcaaagtagatc (SEQ ID NO : 93)

A noter que les nucléotides d'uracile doivent être représentés par le symbole "t" dans le listage des séquences.

En vertu du paragraphe 55 de la norme ST.26, si une séquence de nucléotides contient à la fois des fragments d'ADN et d'ARN, elle doit apparaître sous forme de molécule de type "DNA". La molécule combinée d'ADN et d'ARN doit en outre être décrite 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 de la séquence doit en outre être décrit par la clé de caractérisation "misc_feature", qui indique son emplacement, et par le qualificateur "note", ce dernier précisant s'il s'agit d'un fragment d'ADN ou d'ARN. La séquence divulguée contient deux segments d'ADN (positions 1 à 7 et 43 des nucléotides) et un segment d'ARN (positions 8 à 42 des nucléotides).

Paragraphe(s) pertinent(s) de la norme ST.26 : 7, 14, **55**, 56 et 83

Paragraphe 89 – Clé de caractérisation "CDS"

Exemple 89-1 : Codage de la séquence de nucléotides et séquence d'acides amines codée

Une demande de brevet décrit la séquence nucléotidique ci-après et sa traduction : `atg acc`

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

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

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

La séquence nucléotidique énumérée compte plus de 10 nucléotides définis de manière spécifique.

La séquence d'acides aminés énumérée contient plus de quatre acides aminés définis de manière spécifique.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence nucléotidique doit être présentée sous la forme suivante :

atgaccggaaataaacctgaaaccgatgtttacgaaattttatga (SEQ ID NO : 59)

La séquence nucléotidique doit s'accompagner d'une description supplémentaire utilisant la clé de caractérisation "CDS", et l'élément INSDFeature_location devrait désigner l'ensemble de la séquence, y compris le codon d'arrêt (c'est-à-dire les positions 1 à 45). De plus, le qualificateur "translation" devrait être intégré en prenant la valeur "MTGNKPETDVYEIL". La demande ne divulgue pas le tableau du code génétique qui est appliqué à la traduction (voir annexe 1, section 9, tableau 7). Si le tableau de codes normalisés est appliqué, le qualificateur "transl_table" n'est pas nécessaire; toutefois, si un tableau du code génétique différent est appliqué, il faut indiquer la valeur appropriée figurant dans le tableau 7 pour le qualificateur "transl_table". Enfin, le qualificateur "protein_id" doit être employé pour indiquer le numéro d'identification de la séquence d'acides aminés traduite.

La séquence d'acides aminés.doit être présentée de façon distincte au moyen des codes à une lettre ci-après, et disposer de son propre numéro d'identification de séquence :

MTGNKPETDVYEIL (SEQ ID NO : 60)

Le STOP qui suit la séquence d'acides aminés énumérée ne doit pas être intégré dans la séquence d'acides aminés à l'occasion du listage des séquences.

N.B. La représentation préférée de la séquence indiquée ci-dessus sert à fournir un listage des séquences à la date du dépôt d'une demande de brevet. La même représentation pourra ne pas être applicable à un listage des séquences fourni après cette date, car il faut tenir compte de la question de savoir si l'information fournie pourrait être prise en considération par un office de la propriété intellectuelle pour ajouter des éléments à la divulgation originale.

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.a), 7.b), 26, 28, 89, 90 et 92

Exemple 89-2 : L'emplacement de la caractéristique s'étend au-delà de la séquence divulguée

Une demande de brevet contient la figure suivante, qui révèle une séquence de codage partielle et sa séquence d'acides aminés traduite :

```
cat cac gca gca gaa tgt gga ttt tgt cct caa caa tgg caa gtt cta      48
His His Ala Ala Glu Cys Gly Phe Cys Pro Gln Gln Trp Gln Val Leu
1               5                  10                  15

cgt ggg agt ctg tgc att tgt gag ggt cca gct gaa gga tgg ttc ata      96
Arg Gly Ser Leu Cys Ile Cys Glu Gly Pro Ala Glu Gly Trp Phe Ile
          20                  25                  30

tca aga tgt tgg tta tgg tgt ggg cct caa gtc caa ggc ttt atc ttt     144
Ser Arg Cys Trp Leu Trp Cys Gly Pro Gln Val Gln Gly Phe Ile Phe
      35                  40                  45

gga gaa ggc aag gaa gga ggc ggt gac aga cgg gct gaa gcg agc cct     192
Gly Glu Gly Lys Glu Gly Gly Gly Asp Arg Arg Ala Glu Ala Ser Pro
 50                  55                  60

cag gag ttt tgg gaa tgc act tgg                                    216
Gln Glu Phe Trp Glu Cys Thr Trp
65                  70
```

Figure 1 – Séquence de codage partielle du gène *ITCH1* de l'*Homo sapiens*, qui code les acides aminés 20 à 91 de la protéine ITCH1 comportant 442 acides aminés.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

La demande divulgue une séquence nucléotidique et sa séquence d'acides aminés traduite. Contenant plus de 10 nucléotides définis de manière spécifique, la séquence nucléotidique énumérée doit être intégrée dans un listage des séquences.

La séquence d'acides aminés contient plus de 4 acides aminés définis de manière spécifique et doit donc être intégrée dans un listage des séquences en tant que séquence distincte disposant de son propre numéro d'identification.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence nucléotidique doit être intégrée dans un listage des séquences sous la forme suivante :

catcacgcagcagaatgtggattttgtcctcaacaatggcaagttctacgtgggagtctgtgcatttgtgagggtccagctgaaggatggttcatatcaagatg ttggttatggtgtgggcctcaagtccaaggctttatctttggagaaggcaaggaaggaggcggtgacagacgggctgaagcgagccctcaggagttttggg aatgcacttgg (SEQ ID NO : 94)

La séquence nucléotidique devrait être décrite plus en détail au moyen d'une clé de caractérisation "CDS". L'élément INSDFeature_location doit indiquer l'emplacement de la caractéristique "CDS" dans la séquence, y compris le codon d'arrêt.

La figure décrit une séquence de codage partielle qui ne contient ni le codon de départ ni le codon d'arrêt. Cependant, la description de la séquence indique que le codon de départ se trouve en amont du nucléotide en position 1 et que le codon d'arrêt se trouve en aval de dernier nucléotide en position 216.

La norme ST.26 prescrit que 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. Dès lors, dans l'exemple ci-dessus, le descripteur d'emplacement de la clé de caractérisation "CDS" ne peut comporter de numéros de position au-delà de la série 1 à 216. L'emplacement du codon d'arrêt dans l'élément INSDFeature_location doit être représenté par le symbole ">" pour indiquer que le codon d'arrêt se trouve en aval de la position 216. De même, on peut employer le symbole "<" pour indiquer que le codon de départ se trouve en amont de la position 1. Le descripteur d'emplacement de la clé de caractérisation "CDS" devrait donc apparaître ainsi :

<1..>216

A noter que les symboles "<" et ">" sont des caractères réservés et seront respectivement remplacés par "<" et ">" dans l'instance XML du listage des séquences.

Le qualificateur de "traduction" devrait être intégré en prenant pour valeur la séquence d'acides aminés de la protéine. La figure ne divulgue pas le tableau de codes génétiques s'appliquant à la traduction (voir l'annexe 1, section 9, tableau 7). Si le tableau de codes normalisés est appliqué, le qualificateur "transl_table" n'est pas nécessaire; toutefois, si un tableau du code génétique différent est appliqué, il faut indiquer la valeur appropriée figurant dans le tableau 7 de l'Annexe I pour le qualificateur "transl_table". Enfin, le qualificateur "protein _id" doit être intégré dans la caractéristique "CDS", dont le qualificateur doit indiquer le numéro d'identification de la séquence d'acides aminés traduite.

La séquence d'acides aminés traduite doit apparaître en tant que séquence distincte et disposer de son propre numéro d'identification :

HHAAECGFCPQQWQVLRGSLCICEGPAEGWFISRCWLWCGPQVQGFIFGEGKEGGGDRRAEASPQEFWE CTW (SEQ ID NO : 95)

N.B. La représentation préférée de la séquence indiquée ci-dessus sert à fournir un listage des séquences à la date du dépôt d'une demande de brevet. La même représentation pourra ne pas être applicable à un listage des séquences fourni après cette date, car il faut tenir compte de la question de savoir si l'information fournie pourrait être prise en considération par un office de la propriété intellectuelle pour ajouter des éléments à la divulgation originale.

Paragraphe(s) pertinent(s) de la norme ST.26 : 7, 41, 65, 66, 70, 71, 89 et 92

Paragraphe 92 – Séquence d'acides aminés codée selon une séquence de codage

Exemple 92-1 : Séquence d'acides aminés codée selon une séquence de codage avec introns

Une demande de brevet contient la figure ci-après divulguant une séquence de codage et sa traduction :

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

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

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

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

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

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

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

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

Figure 1 – Les nucléotides apparaissant en gras sont des régions appelées introns.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

La demande divulgue une séquence nucléotidique et son passage à une séquence d'acides aminés (traduction). Contenant plus de 10 nucléotides définis de manière spécifique, la séquence nucléotidique énumérée doit être intégrée dans un listage des séquences comme une séquence unique.

La séquence nucléotidique contient des séquences codantes (exons) séparées par des séquences non codantes (introns). La figure décrit la traduction de la séquence nucléotidique sous la forme de trois séquences d'acides aminés non contiguës. Selon la légende de la figure, les régions de nucléotides apparaissant en gras sont des séquences appelées introns qui, après épissage d'un ARN transcrit, seront traduits en protéines. Il s'ensuit que les trois séquences d'acides aminés sont en fait une séquence énumérée contiguë unique qui contient plus de quatre acides aminés définis de manière spécifique et doit être intégrée dans un listage des séquences comme séquence unique.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence nucléotidique doit être intégrée dans un listage des séquences sous la forme suivante :

atgaagactttcgcagccttgctttccgctgtcactctcgcgctctcggtgcgcgcccaggcggctgtctggagtcaatgtaagtgccgctgcttttcattgatac gagactctacgccgagctgacgtgctaccgtataggtggcggtacaccgggttggacgggcgagaccacttgcgttgctggttcggtttgtacctccttgagc tcagtgagcgactttcaatccgtcgtcattgctcctcatgtattgacgattggccttcatagtcatactctcaatgcgttccgggctccgcaacgtccagcgctcc ggcggccccctcagcgacaacttcaggccccgcacctacggacggaacgtgctcggccagcggggcatggccgccattgacctga (SEQ ID NO : 75)

La séquence nucléotidique doit s'accompagner d'une description supplémentaire mettant en jeu une clé de caractérisation "CDS", et l'élément INSDFeature_location devrait désigner l'emplacement de la séquence codante, y compris le codon d'arrêt indiqué par "Ter". L'élément INSDFeature_location doit utiliser l'opérateur d'emplacement "join" pour indiquer que les produits de traduction codés par les emplacements indiqués sont reliés et forment un polypeptide unique et continu avec le format "join(x1..y1,x2..y2,x3..y3)", par exemple "join (1..79, 142..212, 272..400)". De plus, le qualificateur de "traduction" devrait être intégré en prenant pour valeur la séquence d'acides aminés de la protéine. (On notera que le symbole de fin "Ter" à la dernière position de la séquence ne doit pas être intégré dans la séquence d'acides aminés.) La demande ne divulgue pas le tableau du code génétique qui est appliqué à la traduction (voir annexe 1, section 9, tableau 7). Si le tableau de codes normalisés est appliqué, le qualificateur "transl_table" n'est pas nécessaire; toutefois, si un tableau du code génétique différent est appliqué, il faut indiquer la valeur appropriée figurant dans le tableau 7 pour le qualificateur "transl_table". Enfin, il faut intégrer le qualificateur "protein _id" dont la valeur indiquera le numéro d'identification de séquence de la séquence d'acides aminés traduite. La séquence d'acides aminés doit être intégrée comme une séquence unique :

MKTFAALLSAVTLALSVRAQAAVWSQCGGTPGWTGETTCVAGSVCTSLSSSYSQCVPGSATSSAPAAPSATT SGPAPTDGTCSASGAWPPLT (SEQ ID NO : 76)

Paragraphe(s) pertinent(s) de la norme ST.26 : 7, 26, 28, 57, 67 et 89 à 92

Paragraphe 93 – Séquence primaire et une variante, chacune énumérée par son résidu

Exemple 93-1 : Représentation des variantes énumérées

La description comprend l'alignement de séquence ci-après.

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

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Il est courant dans la branche d'inclure des "points" dans un alignement de séquence pour indiquer que "cette position est la même que la position précédente". On considère que les points dans les séquences de *D.virilis* et de *D.simulans* correspondent à des nucléotides énumérés et définis de manière spécifique; c'est un moyen simple d'indiquer qu'une position donnée est occupée par le même nucléotide que dans la séquence de *D. melanogaster.* De plus, les alignements de séquence présentent souvent le symbole "-" pour indiquer l'absence d'un résidu afin de maximiser l'alignement.

Les séquences de nucléotides de *D. melanogaster* et de *D. simulans* contiennent donc vingt-deux nucléotides énumérés et définis de manière spécifique, tandis que la séquence nucléotidique de *D. virilis* en contient dix-neuf. De ce fait, chaque séquence doit, en vertu du paragraphe 7.a) de la norme ST.26, être intégrée dans un listage des séquences et disposer d'un numéro d'identification de séquence distinct.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence Drosophila melanogaster doit être intégrée dans un listage des séquences sous la forme suivante :

acattgaatctcataccacttt (SEQ ID NO : 61)

La séquence Drosophila virilis doit être intégrée dans un listage des séquences sous la forme suivante :

acatggatcccacgacttt (SEQ ID NO : 62)

La séquence Drosophila simulans doit être intégrée dans un listage des séquences sous la forme suivante :

gtatggcgtcgtatsgtagttt (SEQ ID NO : 63)

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.a), 13 et 93

Exemple 93-2 : Représentation des variantes énumérées

La description comprend le tableau ci-après d'un peptide et de ses variantes fonctionnelles. Dans ce tableau, un espace blanc indique qu'un acide aminé de la variante est le même que l'acide aminé correspondant de la "Séquence" et un "-" indique une suppression de l'acide aminé correspondant de la "Séquence".

Position	1	2	3	4	5	6	7	8	9
Séquence	A	V	L	T	y	L	r	g	E
Variante 1									A
Variante 2			P			P			
Variante 3			A	I	g	y			
Variante 4							-		

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Comme indiqué, un espace blanc dans le tableau indique qu'un acide aminé de la variante est le même que l'acide aminé correspondant de la "Séquence". De ce fait, les acides aminés des séquences variantes sont énumérés et définis de manière spécifique.

Étant donné que les quatre séquences variantes contiennent chacune plus de quatre acides aminés énumérés et définis de manière spécifique, chaque séquence doit, en vertu du paragraphe 7.b) de la norme ST.26, être intégrée dans un listage des séquences et disposer d'un numéro d'identification de séquence distinct.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

AVLTYLRGE (SEQ ID NO : 77)

AVLTYLRGA (SEQ ID NO : 78)

AVPTYPRGE (SEQ ID NO : 79)

AVAIGYRGE (SEQ ID NO : 80)

AVLTYLGE (SEQ ID NO : 81)

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.b), 26 et 93

Exemple 93-3 : Représentation d'une séquence consensus

Une demande de brevet comprend la figure 1 présentant l'alignement séquentiel multiple ci-après.

Consensus	LEG**n**EQFINA**ak**IIRHP**k**Y**nrk**T**ln**NDI**m**LIK
Homo sapiens	LEG**N**EQFINA**AK**IIRHP**Q**Y**DRK**T**LN**NDI**M**LIK
Pongo abelii	LEG**N**EQFINA**AK**IIRHP**Q**Y**DRK**T**VN**NDI**M**LIK
Papio Anubis	LEG**T**EQFINA**AK**IIRHP**D**Y**DRK**T**LN**NDI**L**LIK
Rhinopithecus roxellana	LEG**T**EQFINA**AK**IIRHP**N**Y**NRI**T**LD**NDI**L**LIK
Pan paniscus	LEG**N**EQFINA**AK**IIRHP**K**Y**NRI**T**LN**NDI**M**LIK
Rhinopithecus bieti	LEG**N**EQFINA**TK**IIRHP**K**Y**NGN**T**LN**NDI**M**LIK
Rhinopithecus roxellana	LEG**N**EQFINA**TQ**IIRHP**K**Y**NGN**T**LN**NDI**M**LIK

La séquence consensus comprend des lettres majuscules qui représentent les résidus d'acides aminés conservés, tandis que les lettres minuscules "n", "a", "k", "r", "l" et "m" représentent les résidus d'acides aminés prédominants parmi les séquences alignées.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Les lettres minuscules de la séquence consensus représentent chacune un résidu d'acides aminés unique. En conséquence, la séquence consensus contient, comme chacune des sept autres séquences de la figure 1, au moins quatre acides aminés définis de manière spécifique. Le paragraphe 7.b) de la norme ST.26 prescrit l'intégration des huit séquences dans le listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Dans la séquence consensus, les lettres minuscules sont utilisées comme symboles ambigus pour représenter l'acide aminé prédominant parmi les variantes possibles d'une position particulière. De ce fait, les lettres minuscules "n", "a", "k", "r", "l" et "m" sont des symboles conventionnels utilisés d'une manière non conventionnelle, et la séquence consensus doit être représentée au moyen d'un symbole ambigu à la place de chacune des lettres minuscules.

Il faudrait utiliser le symbole ambigu le plus restrictif. Pour la plupart des positions de la séquence consensus, "X" est le symbole ambigu le plus restrictif; toutefois, le symbole ambigu le plus restrictif pour "D" ou "N" aux positions 20 et 25 est "B". La séquence consensus doit être intégrée dans le listage des séquences sous la forme suivante :

LEGXEQFINAXXIIRHPXYBXXTXBNDIXLIK (SEQ ID NO : 82)

Conformément au paragraphe 27, le symbole "X" sera considéré comme l'équivalent de 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 dans le tableau de caractéristiques. De ce fait, chaque "X" de la séquence consensus doit s'accompagner d'une description supplémentaire dans un tableau de caractéristiques au moyen de la clé de caractérisation "VARIANT" et du qualificateur du type "note" pour indiquer les variantes possibles de chaque position.

Les sept autres séquences doivent être intégrées dans le listage des séquences sous la forme

suivante : LEGNEQFINAAKIIRHPQYDRKTLNNDIMLIK (SEQ ID NO : 83)

LEGNEQFINAAKIIRHPQYDRKTVNNDIMLIK (SEQ ID NO : 84)

LEGTEQFINAAKIIRHPDYDRKTLNNDILLIK (SEQ ID NO : 85)

LEGTEQFINAAKIIRHPNYNRITLDNDILLIK (SEQ ID NO : 86)

LEGNEQFINAAKIIRHPKYNRITLNNDIMLIK (SEQ ID NO : 87)

LEGNEQFINATKIIRHPKYNGNTLNNDIMLIK (SEQ ID NO : 88)

LEGNEQFINATQIIRHPKYNGNTLNNDIMLIK (SEQ ID NO : 89)

N.B. La représentation préférée de la séquence indiquée ci-dessus sert à fournir un listage des séquences à la date du dépôt d'une demande de brevet. La même représentation pourra ne pas être applicable à un listage des séquences fourni après cette date, car il faut tenir compte de la question de savoir si l'information fournie pourrait être prise en considération par un office de la propriété intellectuelle pour ajouter des éléments à la divulgation originale.

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.b), 26, 27, 93 et 97

Paragraphe 94 – Séquence variante divulguée comme une séquence unique avec des résidus alternatifs énumérés

Exemple 94-1 : Représentation d'une séquence unique avec des acides aminés alternatifs énumérés

Une demande de brevet est déposée au titre d'un peptide dont la séquence est la suivante :

i) Gly-Gly-Gly-[Leu ou Ile]-Ala-Thr-[Ser ou Thr]

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

La séquence contenant quatre acides aminés définis de manière spécifique, le paragraphe 7.b) de la norme ST.26 prescrit son intégration dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Le tableau 3 de l'annexe I, section 3 définit le symbole ambigu "J" comme désignant l'isoleucine ou la leucine. De ce fait, la représentation à préférer de la séquence est la suivante :

GGGJATX (SEQ ID NO : 64)

qui doit s'accompagner d'une description supplémentaire dans un tableau de caractéristiques réalisée au moyen de la clé de caractérisation "VARIANT" et du qualificateur du type "note" pour indiquer que "X" est la sérine ou la thréonine.

Une autre solution consiste à représenter la séquence, par exemple, de la façon suivante :

GGGLATS (SEQ ID NO : 65)

qui doit s'accompagner d'une description supplémentaire dans un tableau de caractéristiques réalisée au moyen de la clé de caractérisation "VARIANT" et du qualificateur du type "note" pour indiquer que L peut être remplacé par I, et S par T.

N.B. : La représentation préférée de la séquence indiquée ci-dessus sert à fournir un listage des séquences à la date du dépôt d'une demande de brevet. La même représentation pourra ne pas être applicable à un listage des séquences fourni après cette date, car il faut tenir compte de la question de savoir si l'information fournie pourrait être prise en considération par un office de la propriété intellectuelle pour ajouter des éléments à la divulgation originale.

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.b), 8, 26, 27, 94 et 97

Exemple 94-2 – Représentation d'une séquence unique avec des acides aminés alternatifs énumérés qui peuvent être des acides aminés modifiés

Une demande de brevet décrit le polypeptide suivant :

Leu-Glu-Tyr-Cys-Leu-Lys-Arg-Trp-Xaa-Glu-Thr-Ile-Ser-His-Cys-Ala-Trp

où Xaa peut être Ile, Ala, Phe, Tyr, aIle, MeIle ou Nle.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Le peptide énuméré fournit 16 acides aminés définis avec précision. Par conséquent, la séquence doit être intégrée dans un listage des séquences comme l'exige le paragraphe 7.b) de la norme ST.26.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Le symbole ambigu le plus restrictif pouvant englober "Ile, Ala, Phe, Tyr, aIle, MeIle ou Nle" est "X", la séquence doit être intégrée dans un listage de séquences comme suit :

LEYCLKRWXETISHCAW (SEQ ID NO: 96)

Le paragraphe 30 de la norme ST.26 requiert qu'"[u]n acide aminé modifié doit être accompagné d'une description supplémentaire dans le tableau des caractéristiques". Cependant, il n'exige pas l'utilisation d'une clé de caractérisation spécifique pour décrire les acides aminés modifiés. Même si le paragraphe 30 décrit l'utilisation des clés de caractérisation "CARBOHYD", "LIPID", "MOD_RES" et "SITE", ces clés sont plus appropriées pour les scénarios où l'acide aminé modifié ne figure pas dans la liste d'alternatives d'un emplacement donné. Dans cet exemple, la clé de caractérisation "VARIANT" satisfait à l'exigence du paragraphe 30, puisqu'elle permet de représenter toutes les alternatives à l'emplacement de la variante. Ainsi, la clé "VARIANT" accompagnée du qualificateur "note" "Ile, Ala, Phe, Tyr, aIle, MeIle ou Nle" en tant que valeur de qualificateur doit être utilisée pour décrire l'emplacement de la variante en position 9. L'utilisation d'une deuxième clé de caractérisation telle que "SITE" et du qualificateur "note" peut être employée pour mieux définir les acides aminés modifiés trouvés en position 9.

Paragraphe(s) pertinent(s) de la norme ST.26 : 3.a), 7.b), 26, 27, 30, **94**, 96 et annexe I, section 4, tableau 4

Paragraphe 95.a) – Toute séquence variante divulguée uniquement par référence à une séquence primaire comportant plusieurs variations indépendantes

Exemple 95.a)-1 : Représentation d'une séquence variante par annotation de la séquence primaire

Une demande contient la divulgation ci-après :

"Le fragment peptidique 1 est Gly-Leu-Pro-Xaa-Arg-Ile-Cys, où Xaa peut représenter tout acide aminé…

Dans un autre mode de réalisation, le fragment peptidique 1 est Gly-Leu-Pro-Xaa-Arg-Ile-Cys, où Xaa peut représenter Val, Thr ou Asp….

Dans un autre mode de réalisation, le fragment peptidique 1 est Gly-Leu-Pro-Xaa-Arg-Ile-Cys, où Xaa peut représenter Val."

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Dans chacun des trois modes de réalisation divulgués, le "fragment peptidique 1" contient au moins six acides aminés définis de manière spécifique; de ce fait, la séquence doit être intégrée dans un listage des séquences, comme le prescrit le paragraphe 7.b) de la norme ST.26.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

Dans cet exemple, la séquence énumérée du "fragment peptidique 1" est divulguée trois fois, en tant que trois modes de réalisation différents, chacun donnant lieu à une description différente de Xaa. Dans cet exemple, "X" est le symbole ambigu le plus restrictif pour la position Xaa.

La norme ST.26 prescrit d'intégrer une seule fois la séquence énumérée divulguée. Dans le plus englobant des trois modes de réalisation, Xaa représente tout acide aminé (voir l'Introduction au présent document). De ce fait, la séquence à intégrer dans le listage des séquences est la suivante :

GLPXRIC (SEQ ID NO : 66)

Conformément au paragraphe 27, "X" sera considéré comme l'équivalent de 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 dans le tableau de caractéristiques. Étant donné que dans la séquence SEQ ID NO : 66, "X" représente "tout acide aminé", il doit être annoté au moyen de la clé de caractérisation "VARIANT" et d'un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

Chaque fois que possible, chaque "X" devrait être annoté individuellement. Cependant, une région de résidus "X" contigus ou un grand nombre de résidus "X" dispersés dans l'ensemble de la séquence peuvent être décrits conjointement par la clé de caractérisation "VARIANT" en employant la syntaxe "x..y" pour désigner le descripteur d'emplacement, où x et y sont les positions du premier et du dernier résidu "X", et par un qualificateur "note" prenant la valeur "X peut être tout acide aminé".

Comme indiqué dans l'introduction au présent document, il est fortement conseillé d'intégrer toutes séquences supplémentaires de première importance pour la divulgation ou les revendications de l'invention.

Pour l'exemple ci-dessus, il est vivement conseillé d'intégrer dans le listage des séquences les trois séquences supplémentaires ci-après et d'attribuer à chacune son propre numéro d'identification de séquence :

GLPVRIC (SEQ ID NO : 67)

GLPTRIC (SEQ ID NO : 68)

GLPDRIC (SEQ ID NO : 69)

N.B. La représentation préférée de la séquence indiquée ci-dessus sert à fournir un listage des séquences à la date du dépôt d'une demande de brevet. La même représentation pourra ne pas être applicable à un listage des séquences fourni après cette date, car il faut tenir compte de la question de savoir si l'information fournie pourrait être prise en considération par un office de la propriété intellectuelle pour ajouter des éléments à la divulgation originale.

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.b), 26, 27 et **95.a)**

Paragraphe 95.b) – Toute séquence variante divulguée uniquement par référence à une séquence primaire comportant plusieurs variations interdépendantes

Exemple 95.b)-1 : Représentation des séquences variantes individuelles comportant plusieurs variations interdépendantes

Une demande de brevet décrit la séquence consensus ci-après :

cgaatgn1cccactacgaatgn2cacgaatgn3cccaca

où n1, n2, et n3 peuvent être a, t, g ou c.

Plusieurs séquences variantes sont divulguées comme suit :

si n1 représente a, n2 et n3 représentent t, g ou c;

si n1 représente t, n2 et n3 représentent a, g ou c;

si n1 représente g, n2 et n3 représentent t, a ou c;

si n1 représente c, n2 et n3 représentent t, g ou a.

Question 1 : La norme ST.26 prescrit-elle l'intégration de la ou des séquences?

OUI

Contenant plus de 10 nucléotides énumérés et "définis de manière spécifique", la séquence doit, en vertu du paragraphe 7.a) de la norme ST.26, être intégrée dans un listage des séquences.

Question 3 : Comment la ou les séquences devraient-elles être représentées dans le listage des séquences?

La séquence énumérée contient plus de 10 nucléotides définis de manière spécifique et trois résidus "n". La norme ST.26 prescrit l'intégration de la séquence énumérée divulguée et, lorsqu'il convient d'employer un symbole ambigu, il faut choisir le plus restrictif. Dans cet exemple, n1, n2, et n3 peuvent représenter a, t, g ou c, si bien que "n" est le symbole ambigu le plus restrictif. La séquence à intégrer dans le listage des séquences est donc la suivante :

cgaatgncccactacgaatgncacgaatgncccaca (SEQ ID NO : 70)

Le paragraphe 15 de la norme ST.26 indique que "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 dans le tableau de caractéristiques. Comme la valeur de chaque résidu "n"dans cette séquence est équivalent à la valeur par défaut "a", "c", "g", ou "t", aucune annotation supplémentaire n'est exigée.La séquence énumérée comporte des variations à trois emplacements distincts et les occurrences de ces variations sont interdépendantes. Comme indiqué dans l'introduction au présent document, il est fortement conseillé d'intégrer les séquences supplémentaires qui représentent des modes de réalisation supplémentaires constituant une partie essentielle de l'invention. De ce fait, et conformément au paragraphe 95.b) de la norme ST.26, les modes de réalisation supplémentaires devraient être intégrés dans un listage des séquences en tant que quatre séquences distinctes, chacune disposant de son propre numéro d'identification de séquence :

cgaatgacccactacgaatgbcacgaatgbcccaca (SEQ ID NO : 71)

cgaatgtcccactacgaatgvcacgaatgvcccaca (SEQ ID NO : 72)

cgaatggcccactacgaatghcacgaatghcccaca (SEQ ID NO : 73)

cgaatgccccactacgaatgdcacgaatgdcccaca (SEQ ID NO : 74)

(On notera que b = t, g ou c; v = a, g ou c; h = t, a ou c; et d = t, g ou a; voir annexe I, section 1, tableau 1)

En vertu du paragraphe 15 de la norme ST.26, il faut utiliser le symbole le plus restrictif pour représenter les positions variables. Il s'ensuit que n2 et n3 ne doivent pas être représentés par "n" dans la séquence.

N.B. La représentation préférée de la séquence indiquée ci-dessus sert à fournir un listage des séquences à la date du dépôt d'une demande de brevet. La même représentation pourra ne pas être applicable à un listage des séquences fourni après cette date, car il faut tenir compte de la question de savoir si l'information fournie pourrait être prise en considération par un office de la propriété intellectuelle pour ajouter des éléments à la divulgation originale.

Paragraphe(s) pertinent(s) de la norme ST.26 : 7.a), 15, et 95b)

[L'appendice de l'annexe VI suit]

APPENDICE

DOCUMENT D'ORIENTATION SÉQUENCES EN XML

L'appendice est disponible à l'adresse suivante :
https://www.wipo.int/standards/en/xml_material/st26/st26-annex-vi-appendix-guidance-document-sequences_vi_7.xml

[L'annexe VII suit]

ANNEXE VII

RECOMMANDATION RELATIVE A LA CONVERSION D'UN LISTAGE DES SEQUENCES DE LA NORME ST.25 A LA NORME ST.26 : AJOUT OU SUPPRESSION EVENTUELS D'ELEMENTS

Version 1.7

Révision approuvée au Comité des normes de l'OMPI (CWS) à sa onzième session le 8 décembre 2023

Introduction

Les exigences en matière de présentation de séquences de nucléotides et d'acides aminés diffèrent entre la norme ST.26 et la norme ST.25 de l'OMPI. Il convient donc de déterminer si la norme ST.26 impose d'ajouter ou de supprimer des éléments dans un listage des séquences faisant partie intégrante d'une demande internationale déposée conformément à cette norme, étant entendu que ces éléments pourraient ne pas figurer dans une demande dont la priorité est revendiquée.

Portée du document

Le présent document concerne les exigences énoncées dans la norme ST.26 ainsi que toutes les conséquences qui pourraient en découler. Il ne traite pas individuellement de chaque scénario possible; si les moyens prévus dans la norme ST.26 pour représenter les informations figurant dans un listage des séquences selon la norme ST.25 ne sont pas suffisants, il est toujours possible de faire apparaître ces informations dans la description de la demande afin d'éviter la suppression d'éléments.

Recommandations relatives à l'ajout ou la suppression éventuels d'éléments

L'examen des questions abordées dans le présent document montre que la transformation d'un listage des séquences de la norme ST.25 à la norme ST.26 ne devrait pas nécessiter par elle-même l'ajout ou la suppression d'éléments, surtout si le listage effectué au titre de la norme ST.25 était entièrement conforme à celle-ci. Néanmoins, dans certains cas le déposant devra être prudent. Des recommandations ont été établies pour éviter que des éléments ne soient ajoutés ou supprimés.

Scénario 1

La norme ST.25 prévoit l'emploi d'identifiants numériques pour étiqueter différents types de données, par exemple <110> pour le nom du déposant. La norme ST.26 prévoit pour sa part l'emploi de termes anglais pour nommer et décrire les données.

Recommandation

Les termes prévus dans la norme ST.26 ne font que décrire le contenu des données; dès lors, l'emploi de ces noms et descripteurs ne constitue pas un ajout d'éléments.

Scénario 2

La norme ST.26 fait obligation d'indiquer : a) les séquences ramifiées; b) les séquences comportant des acides aminés D; c) les analogues nucléotidiques; et d) les séquences comportant des sites abasiques. L'obligation d'indiquer ces séquences ou leur interdiction n'est pas clairement établie dans la norme ST.25.

Recommandation

Les informations divulguées dans la demande devraient être suffisantes pour permettre de représenter ces séquences dans un listage des séquences conforme à la norme ST.26, alors qu'elles n'auraient peut-être pas été indiquées dans un listage conforme à la norme ST.25. Pour certains types d'informations exigés par la norme ST.26, il faut veiller à ne pas ajouter d'éléments au-delà de ce qui a déjà été divulgué (voir par exemple l'analyse du scénario 4 concernant le qualificateur "mol_type" destiné aux séquences de nucléotides).

Scénario 3

La norme ST.26 interdit les séquences comportant moins de 10 nucléotides définis de manière spécifique ("n" exclu) et moins de quatre acides aminés définis de manière spécifique ("X" exclu).

Recommandation

Les séquences interdites peuvent être ajoutées dans le corps de la demande si elles n'y figurent pas déjà.

Scénario 4

Pour les séquences de nucléotides comme pour les séquences d'acides aminés, la norme ST.26 impose d'employer les clés de caractérisation "source" avec deux qualificateurs obligatoires, l'un d'eux étant 'mol_type'. La norme ST.25 comporte une clé de caractérisation correspondante pour les séquences de nucléotides (qui est rarement employée) sans les qualificateurs correspondants, et ne comporte aucune clé de caractérisation correspondante pour les séquences d'acides aminés.

Séquences de nucléotides

ST.26 - clé de caractérisation 5.37 "source"; qualificateur 6.39 "mol_type" (voir le paragraphe 75 de la norme ST.26)

Qualificateur	Valeur
mol_type	genomic DNA
	genomic RNA
	mRNA
	tRNA
	rRNA
	other DNA (applies to synthetic molecules)
	other RNA (applies to synthetic molecules)
	transcribed RNA
	viral cRNA
	unassigned DNA (applies where *in vivo* molecule is unknown)
	unassigned RNA (applies where *in vivo* molecule is unknown)

Séquences d'acides aminés

ST.26 - clé de caractérisation 7.30 "source"; qualificateur 8.1 "mol_type" (voir le paragraphe 75 de la norme ST.26)

Qualificateur	Valeur
MOL_TYPEmol_type	protein

Recommandation

La seule difficulté tient aux valeurs du vocabulaire contrôlé liées au qualificateur "mol_type" dans les séquences de nucléotides. Certaines des valeurs indiquées plus haut peuvent ne pas être entièrement compatibles avec la divulgation. Néanmoins, on peut éviter d'ajouter des éléments si l'on emploie la valeur la plus générique d'une séquence particulière, par exemple "other DNA" et "other RNA" pour une molécule synthétique, ou encore "unassigned DNA" et "unassigned RNA" pour une molécule *in vivo*.

Scénario 5

Lorsqu'une séquence comporte le symbole "Xaa", la norme ST.25 prévoit l'ajout d'autres informations concernant ce résidu dans le champ <223>, qui est annexé aux champs <221> (Nom/clé) et <222> (Emplacement). Cette norme ne définit pas de valeur par défaut pour le symbole "Xaa" (appelé "X" dans la norme ST.26). En revanche, la norme ST.26 établit cette valeur par défaut; c'est pourquoi il n'est pas toujours nécessaire de fournir des informations supplémentaires. Les annotations "tout acide aminé" et "tout acide aminé naturel" renseignent très souvent les variables "Xaa" ou "X" dans les séquences de peptides. On pourrait considérer que ces termes recouvrent des acides aminés différents de ceux qui sont énumérés dans les tableaux d'acides aminés figurant dans les normes ST.25 ou ST.26. Dans la norme ST.26, le symbole "X" sans autre annotation prend par défaut la valeur de n'importe quel acide aminé parmi les 22 qui sont énumérés dans l'annexe I (voir le tableau 3 de la section 3). Cette valeur par défaut peut en elle-même constituer un élément ajouté ou supprimé, et avoir par conséquent un effet néfaste sur la portée de la demande de brevet lors de la conversion de la norme ST.25 à la norme ST.26.

Recommandations

a) Si le listage des séquences selon la norme ST.25 comprend des champs <221> (Nom/clé), <222> (Emplacement correspondant au symbole "Xaa") et <223> (Autres informations sur le symbole "Xaa"), et si le champ <221> (Nom/clé) constitue aussi une clé de caractérisation adéquate selon la norme ST.26, en correspondant par exemple aux clés "SITE", "VARIANT" ou "UNSURE", il convient d'utiliser la clé de caractérisation de la norme ST.26. En outre, pour éviter toute suppression éventuelle d'éléments, les informations indiquées dans le champ <223> doivent aussi apparaître dans le qualificateur "note" annexé au listage.

b) Si le listage des séquences selon la norme ST.25 comprend des champs <221> (Nom/clé), <222> (Emplacement correspondant au symbole "Xaa") et <223> (Autres informations sur le symbole "Xaa"), et si le champ <221> (Nom/clé) ne constitue pas de clé de caractérisation selon la norme ST.26, il convient d'utiliser la clé de caractérisation "SITE" ou "REGION" de la norme ST.26, selon les besoins. En outre, pour éviter toute suppression éventuelle d'éléments, les informations indiquées dans le champ <223> et le nom inadéquat du champ <221> doivent aussi apparaître dans le qualificateur "note" annexé au listage. Ainsi, tout listage selon la norme ST.25 employant un nom qui n'est pas prévu dans la norme ST.25 ou ST.26, par exemple <221> (Variable), et s'accompagnant d'autres informations dans le champ <223> (Xaa) décrit un acide aminé. Dans le présent exemple, la valeur du qualificateur "note" selon la norme ST.26 serait la suivante : "Variable – Xaa est un acide aminé".

c) Si le listage des séquences selon la norme ST.25 ne contient aucun champ <221>, <222>, ou <223> correspondant au symbole "Xaa", ou si des champs <221> et <222> correspondant au symbole "Xaa" apparaissent mais qu'aucune information ne figure dans un champ <223> correspondant (aucun de ces deux scénarios n'est conforme à la norme ST.25, mais ils se sont néanmoins produits), toute information indiquée dans le corps de la demande pour renseigner le symbole "Xaa" devrait apparaître dans le qualificateur "note" conforme à la norme ST.26 et s'accompagner d'une clé de caractérisation adéquate, par exemple "SITE", "REGION" ou "UNSURE", ainsi que d'une indication de l'emplacement.

Scénario 6

Dans la norme ST.25, l'uracile est représenté dans la séquence par le symbole "u" et la thymine par le symbole "t". Dans la norme ST.26, l'uracile et la thymine sont tous deux représentés dans la séquence par le symbole "t" sans autre annotation, le symbole "t" représentant l'uracile dans l'ARN et la thymine dans l'ADN.

Recommandations

a) Lorsqu'une séquence d'ADN contient de l'uracile, celui-ci est considéré, dans le contexte de la norme ST.26, comme un nucléotide modifié. Dès lors, la norme ST.26 fait obligation de représenter l'uracile par un "t", de lui adjoindre une description supplémentaire dans la clé de caractérisation "modified_base", et de donner au qualificateur "mod_base" la valeur "OTHER" et au qualificateur "note" la valeur "uracil". Cette annotation selon la norme ST.26 n'est pas considérée comme un élément ajouté si la séquence d'ADN décrite selon la norme ST.25 contenait un symbole "u".

b) Lorsqu'une séquence d'ARN contient de la thymine, celle-ci est considérée, dans le contexte de la norme ST.26, comme un nucléotide modifié. Dès lors, la norme ST.26 fait obligation de représenter la thymine par un "t", de lui adjoindre une description supplémentaire dans la clé de caractérisation "modified_base", et de donner au qualificateur "mod_base" la valeur "OTHER" et au qualificateur "note" la valeur "thymine". Cette annotation selon la norme ST.26 n'est pas considérée comme un élément ajouté si la séquence d'ARN décrite selon la norme ST.25 contenait un symbole "t".

Scénario 7

Les normes ST.25 et ST.26 prévoient toutes deux que des nucléotides modifiés ou des acides aminés doivent être accompagnés d'une description supplémentaire. Dans la norme ST.26, l'identité d'un nucléotide modifié peut être indiquée, le cas échéant, par l'une des abréviations figurant dans le tableau 2 de la section 2 de l'annexe I; si tel n'est pas le cas, le nom complet non abrégé du nucléotide modifié doit être indiqué. De même, l'identité d'un acide aminé modifié peut être indiquée, le cas échéant, par l'une des abréviations figurant dans le tableau 4 de la section 4 de l'annexe I; si tel n'est pas le cas, le nom complet non abrégé de l'acide aminé modifié doit être indiqué. Inversement, si un résidu modifié ne figure pas dans un tableau de la norme ST.25, il n'est pas obligatoire d'indiquer son nom complet non abrégé; il n'est pas rare qu'une abréviation soit alors employée.

Recommandations

a) Lorsque seul un nom abrégé n'apparaissant ni dans le tableau 2 de la section 2 de l'annexe I ni dans le tableau 4 de la section 4 de l'annexe I a été employé à la fois dans la demande et dans un listage des séquences selon la norme ST.25 pour un nucléotide modifié ou un acide aminé modifié, et que l'homme du métier sait que ce nom abrégé ne désigne qu'un seul nucléotide modifié ou acide aminé modifié bien précis, l'emploi du nom complet non abrégé ne constitue pas en lui-même un élément ajouté.

b) Lorsque seul un nom abrégé n'apparaissant ni dans le tableau 2 de la section 2 de l'annexe I ni dans le tableau 4 de la section 4 de l'annexe I a été employé à la fois dans la demande et dans un listage des séquences selon la norme ST.25 pour un nucléotide modifié ou un acide aminé modifié (et que la demande ne comporte aucune structure chimique), et que l'homme du métier ne sait pas que ce nom abrégé ne désigne qu'un seul nucléotide modifié ou acide aminé modifié bien précis, soit parce que l'abréviation lui est totalement inconnue, soit parce qu'elle pourrait représenter

différents nucléotides modifiés ou acides aminés modifiés, il est impossible de se conformer à la norme ST.26 sans ajouter un élément. Bien entendu, dans ce cas particulier la demande établissant une priorité et le listage des séquences sont eux-mêmes vagues. Pour éviter une éventuelle suppression d'éléments, le nom abrégé figurant dans le listage des séquences selon la norme ST.25 devrait être placé dans un qualificateur "note" conforme à la norme ST.26 en plus du nom complet non abrégé du nucléotide modifié ou de l'acide aminé modifié. Ce nom complet non abrégé, qui est prévu dans les listages des séquences définis par la norme ST.26, n'aura pas la priorité par rapport à la demande antérieure. Il faut veiller à faire apparaître le nom non abrégé dans le listage des séquences original (selon la norme ST.25) et dans la divulgation de la demande afin d'éviter des problèmes par la suite.

Scénario 8

La norme ST.25 comporte un certain nombre de clés de caractérisation qui n'existent pas dans la norme ST.26. Les déposants doivent donc veiller à reprendre les informations figurant dans ces clés de caractérisation d'une manière conforme à la norme ST.26 sans avoir à ajouter ou à supprimer d'éléments.

Recommandations

Le tableau suivant fournit des orientations quant à la manière de reprendre les informations figurant dans une ancienne clé de caractérisation de la norme ST.25 en respectant la norme ST.26 et sans avoir à ajouter ou à supprimer d'éléments. Les numéros 1 à 23 correspondent à des clés de caractérisation de séquences de nucléotides, et les numéros 24 à 43 correspondent à des clés de séquences d'acides aminés.

N°	Clé de caractérisation du champ <221> dans la norme ST.25	Équivalent dans la norme ST.26		
		Clé de caractérisation	Qualificateur	Valeur du qualificateur
1	allele	misc_feature	Allele	valeur du champ <223>
2	attenuator	regulatory[3]	regulatory_class[3]	"attenuator"
			note *(if <223> present)*	valeur du champ <223>
3	CAAT_signal	regulatory[3]	regulatory_class[3]	"CAAT_signal"
			note *(if <223> present)*	valeur du champ <223>
4	conflict	misc_feature	Note	"conflict" et valeur du champ <223>
5	enhancer	regulatory[4]	regulatory_class[3]	"enhancer"
			note *(if <223> present)*	valeur du champ <223>
6	GC_signal	regulatory[3]	regulatory_class[3]	"GC_signal"
			note *(if <223> present)*	valeur du champ <223>
7	LTR	mobile_element[3]	rpt_type[3]	"long_terminal_repeat"
			note *(if <223> present)*	valeur du champ <223>
8	misc_signal	regulatory[3]	regulatory_class[3]	"other"
			note *(if <223> present)*	valeur du champ <223>
9	mutation	variation	Note	"mutation" et valeur du champ <223>
10	old_sequence	misc_feature	Note	"old_sequence" et valeur du champ <223>
11	polyA_signal	regulatory[3]	regulatory_class[3]	"polyA_signal_sequence"
			note *(if <223> present)*	valeur du champ <223>
12	promoter	regulatory[3]	regulatory_class[3]	"promoter"
			note *(if <223> present)*	valeur du champ <223>
13	RBS	regulatory[3]	regulatory_class[3]	"ribosome_binding_site"
			note *(if <223> present)*	valeur du champ <223>
14	repeat_unit a) when repeat_region not used	misc_feature	Note	"repeat_unit" et valeur du champ <223>
	repeat_unit b) when repeat_region used	repeat_region	rpt_unit_range	1st residue..last residue
			note *(if <223> present)*	valeur du champ <223>
15	satellite	repeat_region	Satellite	"satellite" (ou "microsatellite" ou "minisatellite" – si pris en charge)
			note *(if <223> present)*	valeur du champ <223>
16	scRNA	ncRNA[3]	ncRNA_class[3]	"scRNA"
			note *(if <223> present)*	valeur du champ <223>
17	snRNA	ncRNA[3]	ncRNA_class[3]	"snRNA"
			note *(if <223> present)*	valeur du champ <223>
18	TATA_signal	regulatory[3]	regulatory_class[3]	"TATA_box"[5]
19	terminator	regulatory[3]	regulatory_class[3]	"terminator"
20	3'clip	misc_feature	Note	"3'clip" et valeur du champ <223>
21	5'clip	misc_feature	Note	"5'clip" et valeur du champ <223>
22	-10_signal	regulatory[3]	regulatory_class[3]	"minus_35_signal"
23	-35_signal	regulatory[3]	regulatory_class[3]	"minus_35_signal"
			note *(if <223> present)*	valeur du champ <223>

[4] La norme ST.26 peut prévoir de remplacer une clé de caractérisation de la norme ST.25, par exemple TATA_signal, par une clé de caractérisation, un qualificateur ou une valeur plus généraux, par exemple regulatory/regulatory_class/ TATA_box.

[5] Dans un tel cas, afin d'éviter l'ajout d'un objet pouvant entraîner une perte partielle de priorité, il est recommandé d'inclure le terme plus limité "TATA_signal" dans un qualificateur "note" comme indiqué dans le tableau ci dessus (point n° 18). Si, dans de rares cas, le déposant considère que l'utilisation de la valeur "TATA_box" pour le qualificateur "classe réglementaire" n'est pas appropriée, la valeur "other" peut être utilisée à la place de "TATA_box". Dans ce cas, le terme "TATA_signal" doit être inclus dans un qualificateur "note" associé à la clé de caractérisation "réglementaire".

N°	Clé de caractérisation du champ <221> dans la norme ST.25	Équivalent dans la norme ST.26		
		Clé de caractérisation	**Qualificateur**	**Valeur du qualificateur**
24	NON_CONS	Cette caractéristique décrit une brèche composée d'un nombre inconnu de résidus dans une même séquence, ce qui est interdit aussi bien par la norme ST.25 (paragraphe 22) que par la norme ST.26 (paragraphe 37). Dès lors, il convient d'intégrer chaque région composée de résidus définis de manière spécifique et visée par le paragraphe 7 de la norme ST.26 dans le listage des séquences à titre de séquence distincte, et de lui attribuer son propre numéro d'identification de séquence. Pour éviter d'ajouter ou de supprimer un élément, chacune de ces séquences doit être annotée pour indiquer qu'elle fait partie d'une séquence plus longue comportant une brèche indéfinie.		
		REGION	note	Description
		Description de l'emplacement et de la cible de la liaison de la séquence, par exemple "Ce résidu est relié par une liaison N terminale à un peptide comportant un Gly Gly N terminal et une brèche de longueur indéfinie".		
25	SIMILAR	REGION	note	"SIMILAR" et valeur du champ <223> si présent
26	THIOETH	CROSSLNK	note	"THIOETH" et valeur du champ <223> si présent
		Pour plus de détails sur les informations d'emplacement, voir le commentaire sur la clé de caractérisation CROSSLNK dans l'annexe I de la norme ST.26.		
27	THIOLEST	CROSSLNK	note	"THIOLEST" et valeur du champ <223> si présent
		Pour plus de détails sur les informations d'emplacement, voir le commentaire sur la clé de caractérisation CROSSLNK dans l'annexe I de la norme ST.26.		
28	VARSPLIC	Analysé dans le scénario 13		
29	ACETYLATION	MOD_RES	note	"ACETYLATION" et valeur du champ <223> si présent
			note	Informations requises si possible en vertu du commentaire sur la clé de caractérisation "MOD_RES" figurant dans l'annexe I de la norme ST.26 (sans ajouter d'élément)
30	AMIDATION	MOD_RES	note	"AMIDATION" et valeur du champ <223> si présent
			note	Informations requises si possible en vertu du commentaire sur la clé de caractérisation "MOD_RES" figurant dans l'annexe I de la norme ST.26 (sans ajouter d'élément)
31	BLOCKED	MOD_RES	note	"BLOCKED" et valeur du champ <223> si présent
			note	Informations requises si possible en vertu du commentaire sur la clé de caractérisation "MOD_RES" figurant dans l'annexe I de la norme ST.26 (sans ajouter d'élément)
32	FORMYLATION	MOD_RES	note	"FORMYLATION" et valeur du champ <223> si présent
33	GAMMA-CARBOXYGLUTAMIC ACID HYDROXYLATION	MOD_RES	note	"GAMMA-CARBOXYLGLUTAMIC ACID HYDROXYLATION" et valeur du champ <223> si présent
			note	Informations requises si possible en vertu du commentaire sur la clé de caractérisation "MOD_RES" figurant dans l'annexe I de la norme ST.26 (sans ajouter d'élément)
34	METHYLATION	MOD_RES	note	"METHYLATION" et valeur du champ <223> si présent
			note	Informations requises si possible en vertu du commentaire sur la clé de caractérisation "MOD_RES" figurant dans l'annexe I de la norme ST.26 (sans ajouter d'élément)

N°	Clé de caractérisation du champ <221> dans la norme ST.25	Équivalent dans la norme ST.26		
		Clé de caractérisation	Qualificateur	Valeur du qualificateur
35	PHOSPHORYLATION	MOD_RES	note	"PHOSPHORYLATION" et valeur du champ <223> si présent
			note	Informations requises si possible en vertu du commentaire sur la clé de caractérisation "MOD_RES" figurant dans l'annexe I de la norme ST.26 (sans ajouter d'élément)
36	PYRROLIDONE CARBOXYLIC ACID	MOD_RES	note	"PYRROLIDONE CARBOXYLIC ACID" et valeur du champ <223> si présent
			note	Informations requises si possible en vertu du commentaire sur la clé de caractérisation "MOD_RES" figurant dans l'annexe I de la norme ST.26 (sans ajouter d'élément)
37	SULFATATION	MOD_RES	note	"SULFATATION" et valeur du champ <223> si présent
			note	Informations requises si possible en vertu du commentaire sur la clé de caractérisation "MOD_RES" figurant dans l'annexe I de la norme ST.26 (sans ajouter d'élément)
38	MYRISTATE	LIPID	note	"MYRISTATE" et valeur du champ <223> si présent
			note	Informations requises si possible en vertu du commentaire sur la clé de caractérisation "LIPID" figurant dans l'annexe I de la norme ST.26 (sans ajouter d'élément)
39	PALMITATE	LIPID	note	"PALMITATE" et valeur du champ <223> si présent
			note	Informations requises si possible en vertu du commentaire sur la clé de caractérisation "LIPID" figurant dans l'annexe I de la norme ST.26 (sans ajouter d'élément)
40	FARNESYL	LIPID	note	"FARNESYL" et valeur du champ <223> si présent
			note	Informations requises si possible en vertu du commentaire sur la clé de caractérisation "LIPID" figurant dans l'annexe I de la norme ST.26 (sans ajouter d'élément)

N°	Clé de caractérisation du champ <221> dans la norme ST.25	Équivalent dans la norme ST.26		
		Clé de caractérisation	Qualificateur	Valeur du qualificateur
41	GERANYL-GERANYL	LIPID	note	"GERANYL-GERANYL" et valeur du champ <223> si présent
			note	Informations requises si possible en vertu du commentaire sur la clé de caractérisation "LIPID" figurant dans l'annexe I de la norme ST.26 (sans ajouter d'élément)
42	GPI-ANCHOR	LIPID	note	"GPI-ANCHOR" et valeur du champ <223> si présent
			note	Informations requises si possible en vertu du commentaire sur la clé de caractérisation "LIPID" figurant dans l'annexe I de la norme ST.26 (sans ajouter d'élément)
43	N-ACYL DIGLYCERIDE	LIPID	note	"N-ACYL DIGLYCERIDE" et valeur du champ <223> si présent
			NOTE	Informations requises si possible en vertu du commentaire sur la clé de caractérisation "LIPID" figurant dans l'annexe I de la norme ST.26 (sans ajouter d'élément)

Scénario 9

Certaines clés de caractérisation présentes à la fois dans les normes ST.25 et ST.26, tant pour les séquences de nucléotides que pour les séquences d'acides aminés, ont des qualificateurs obligatoires dans la norme ST.26, comme indiqué plus loin. La clé de caractérisation "modified_base" pour la séquence de nucléotides est également présente à la fois dans la norme ST.25 et dans la norme ST.26; cependant, le scénario 7 contient des recommandations appropriées. La norme ST.25 ne prévoit pas de qualificateur, mais elle comporte un champ de texte libre portant le numéro <223>. Si les informations figurant dans le champ <223> de la norme ST.25 sont compatibles avec la valeur du qualificateur obligatoire de la norme ST.26, elles doivent être reprises telles quelles. Si le champ <223> de la norme ST.25 est absent ou contient des informations qui ne sont pas compatibles avec la valeur du qualificateur obligatoire de la norme ST.26, le déposant doit veiller à reprendre les informations figurant dans la clé de caractérisation ou dans le champ <223> de la norme ST.25 d'une manière qui soit conforme à la norme ST.26 sans ajouter ni supprimer d'élément.

Séquences de nucléotides[6]

Clé de caractérisation	Qualificateur obligatoire
5.12 – misc_binding	6.3 – bound_moiety
5.30 – protein_bind	6.3 – bound_moiety

Recommandations

a) Si le champ <223> de la norme ST.25 est absent ou incompatible et que la description de la demande révèle le nom de la molécule ou du complexe susceptible d'assurer la liaison avec l'emplacement de la caractéristique de l'acide nucléique, ce nom doit être repris dans le qualificateur "bound_moiety".

i) Toute information figurant dans le champ <223> de la norme ST.25 qui est incompatible avec le qualificateur "bound_moiety" doit être reprise dans un qualificateur facultatif adéquat de la clé de caractérisation, par exemple "note".

b) Si le champ <223> de la norme ST.25 est absent ou incompatible et que la description de la demande révèle le nom de la molécule ou du complexe susceptible d'assurer la liaison avec l'emplacement de la caractéristique de l'acide nucléique, il convient d'employer la clé de caractérisation "misc_feature" de la norme ST.26 au lieu de la clé "misc_binding" ou "protein_bind", ainsi que le qualificateur "note".

i) Si le champ <223> de la norme ST.25 est absent, la valeur du qualificateur "note" doit être le nom de la clé de caractérisation de cette norme;

ii) Si le champ <223> de la norme ST.25 contient des informations incompatibles, la valeur du qualificateur "note" doit être le nom de la clé de caractérisation de cette norme auquel s'ajoutent les informations du champ <223>.

[6] Les numéros de référence indiqués dans le tableau ci-après renvoient aux numéros de clé de caractérisation et de qualificateur figurant dans le vocabulaire contrôlé de l'annexe I de la norme ST.26.

Séquences d'acides aminés[4]

Clé de caractérisation	Qualificateur obligatoire
7.2 – BINDING	8.2 – note
7.4 – CARBOHYD	8.2 – note
7.10 – DISULFID	8.2 – note
7.11 – DNA_BIND	8.2 – note
7.12 – DOMAIN	8.2 – note
7.16 – LIPID	8.2 – note
7.17 – METAL	8.2 – note
7.18 – MOD_RES	8.2 – note
7.23 – NP_BIND	8.2 – note
7.29 – SITE	8.2 – note
7.39 – ZN_FING	8.2 – note

Recommandations

a) Si le champ <223> de la norme ST.25 est absent ou incompatible et que la description de la demande révèle les informations spécifiques devant figurer dans le qualificateur obligatoire, ces informations doivent être reprises dans le qualificateur obligatoire "note".

i) Toute information figurant dans le champ <223> de la norme ST.25 qui est incompatible avec le qualificateur obligatoire "NOTE" (voir la définition et les commentaires de la clé de caractérisation) doit être reprise dans un second qualificateur "note".

b) Si le champ <223> de la norme ST.25 est absent ou incompatible et que la description de la demande ne révèle pas les informations spécifiques devant figurer dans le qualificateur obligatoire, il convient d'employer plutôt la clé de caractérisation "SITE" de la norme ST.26 (pour un seul acide aminé) ou "REGION" (pour une série d'acides aminés) ainsi que le qualificateur "note".

i) Si le champ <223> de la norme ST.25 est absent, la valeur du qualificateur "note" doit être le nom de la clé de caractérisation de cette norme;

ii) Si le champ <223> de la norme ST.25 contient des informations incompatibles, la valeur du qualificateur "note" doit être le nom de la clé de caractérisation de cette norme auquel s'ajoutent les informations du champ <223>.

Scénario 10

Chaque clé de caractérisation spécifique de la norme ST.25 comporte un champ <222> permettant d'indiquer un emplacement; cependant, d'une part cette norme ne fait pas obligation d'indiquer l'emplacement pour la plupart des caractéristiques, et d'autre part le format de ces informations d'emplacement n'est pas normalisé. Au demeurant, la norme ST.25 ne comporte pas d'opérateur d'emplacement tel que "join". La norme ST.26 comporte quant à elle des descripteurs et des opérateurs d'emplacement, chacune de ses caractéristiques devant comporter au moins un descripteur d'emplacement. (Les caractéristiques CDS constituent un cas particulier examiné dans le scénario 11).

Recommandations

a) Si le listage des séquences selon la norme ST.25 comporte un champ <222>, l'importation directe et la conversion au format ST.26 ne devraient pas nécessiter l'ajout d'éléments;

b) Si le listage des séquences selon la norme ST.25 ne comporte pas de champ <222>, mais que les informations sur l'emplacement figurent dans la description de la demande, l'importation directe et la conversion au format ST.26 ne devraient pas nécessiter l'ajout d'éléments;

c) Si ni le listage des séquences selon la norme ST.25 ni la description de la demande ne contiennent d'informations sur l'emplacement, la caractéristique s'applique par hypothèse à l'ensemble de la séquence. (Le fait d'indiquer un emplacement moins long que l'ensemble de la séquence sans le justifier dans la description de la demande constituerait probablement un élément ajouté ou supprimé.) Il faut veiller à faire apparaître, dans la mesure du possible, les informations d'emplacement dans le listage des séquences original (selon la norme ST.25) et dans la divulgation de la demande afin d'éviter des problèmes par la suite.

Scénario 11

Selon la norme ST.25, toute séquence codant un polypeptide unique et continu mais ayant été interrompue par une ou plusieurs séquences non codantes (par exemple des introns) apparaît sous la forme de plusieurs caractéristiques CDS, de la manière suivante :

<220>
<221> CDS
<222> (1)..(571)

<220>
<221> CDS
<222> (639)..(859)

La norme ST.26, pour sa part, dispose d'un opérateur d'emplacement "join" qui permet de préciser que les polypeptides codés par les emplacements indiqués sont joints et forment un polypeptide unique et continu. (Note : les normes ST.25 et ST.26 imposent toutes deux que le codon d'arrêt figure dans l'emplacement de la caractéristique CDS.)

Recommandations

a) Si le listage des séquences selon la norme ST.25 ou la description de la demande indique clairement que les séquences de polypeptides codées par plusieurs caractéristiques CDS distinctes constituent un polypeptide unique et continu, toute séquence de codage interrompue par un intron dans une caractéristique CDS particulière doit être représentée par l'opérateur d'emplacement "join", de la manière décrite ci-après, afin de n'ajouter ou de ne supprimer aucun élément :

<INSDFeature_key>CDS</INSDFeature_key>
<INSDFeature_location>join(1..571,639..859)</INSDFeature_location>

b) Si le listage des séquences selon la norme ST.25 ou la description de la demande n'indique pas que les séquences de polypeptides codées par les deux caractéristiques CDS distinctes constituent un polypeptide unique et continu, l'emploi de l'opérateur d'emplacement "join" constituera probablement un ajout d'élément.

Scénario 12

La norme ST.25 dispose que le nom des caractéristiques doit être repris des tableaux 5 ou 6. Toutefois, selon la réglementation des États-Unis d'Amérique, il s'agit d'une recommandation et non d'une obligation. Dès lors, toute séquence d'un listage établi selon la norme ST.25 (et conforme à la réglementation des États-Unis d'Amérique) peut comporter une clé de caractérisation dont le nom est "Custom" et qui n'a pas de correspondance dans la norme ST.26. Il est également possible qu'aucun nom de caractéristique n'ait été indiqué dans le champ <221>, ou que ce champ soit absent. Ces autres scénarios peuvent être gérés de la même manière.

Recommandation

Le nom de la clé de caractérisation "Custom" de la norme ST.25 peut apparaître dans un listage des séquences selon la norme ST.26 sans ajouter d'élément, de la manière suivante :

Type	**Clé de caractérisation du champ <221> dans la norme ST.25**	**Équivalent possible dans la norme ST.26**		
		Clé de caractérisation	**Qualificateur**	**Valeur du qualificateur**
NA	Clé de caractérisation "Custom"	misc_feature	note	Nom de la clé de caractérisation "Custom" et valeur du champ <223> si présent
AA	Clé de caractérisation "Custom"	SITE ou REGION	note	Nom de la clé de caractérisation "Custom" et valeur du champ <223> si présent

Scénario 13

La norme ST.25 comporte la clé de caractérisation "VARSPLIC" qui est définie comme une "description des variants de la séquence produits par un épissage alternatif". Dans la norme ST.26, la clé "VARSPLIC" a été remplacée par une clé de caractérisation plus générale appelée "VAR_SEQ" et définie comme une "description des variants de la séquence produits par un épissage alternatif, l'usage de promoteurs alternatifs, une initiation alternative et un déphasage ribosomique". Il ne faut donc pas utiliser la clé "VAR_SEQ" pour remplacer la clé "VARSPLIC" sans ajouter d'explication supplémentaire dans un listage des séquences conforme à la norme ST.26.

Recommandation

Selon la norme ST.26, la caractéristique "VAR_SEQ" doit être employée conjointement avec le qualificateur "note", dont la valeur doit indiquer que la portée était plus restreinte dans la norme ST.25, par exemple avec la mention "Variant de la séquence produit par un épissage alternatif". Toute information supplémentaire figurant dans un champ <223> de la norme ST.25 devrait aussi être reprise dans le qualificateur "note".

Scénario 14

Si la source de la séquence est artificielle, il convient de porter la mention "Artificial Sequence" dans le champ <213> (Organisme) de la norme ST.25. Selon la norme ST.26, la clé de caractérisation "source" doit comporter le qualificateur "organism", dont la valeur doit être "synthetic construct" et non "Artificial Sequence".

Recommandation

La valeur du qualificateur "organism" de la norme ST.26 doit être "synthetic construct". Pour éviter une éventuelle suppression d'élément, toute explication figurant dans le champ obligatoire <223> de la norme ST.25 doit apparaître dans un qualificateur "note" (de la clé de caractérisation "source").

Scénario 15

Si le nom scientifique de l'organisme source d'une séquence est inconnu, il faut porter la mention "Unknown" dans le champ <213> (Organisme) de la norme ST.25. Selon la norme ST.26, la clé de caractérisation "source" doit comporter le qualificateur "organism", dont la valeur doit être "unidentified" et non "Unknown".

Recommandation

La valeur du qualificateur "organism" selon la norme ST.26 doit être "unidentified". Pour éviter une éventuelle suppression d'élément, toute explication figurant dans le champ obligatoire <223> de la norme ST.25 doit apparaître dans un qualificateur "note" (de la clé de caractérisation "source").

Scénario 16

Pour les acides aminés précédant la protéine mature, par exemple les préséquences, les proséquences et les pré-proséquences ainsi que les séquences signal, la norme ST.25 permet d'ajouter à la liste des acides aminés, à titre facultatif, des numéros négatifs en comptant à rebours et en commençant par l'acide aminé voisin de l'acide portant le numéro 1. La norme ST.26 ne permet pas d'utiliser des numéros négatifs dans l'emplacement de la caractéristique.

Recommandations

a) Si le listage des séquences selon la norme ST.25 comporte une ou plusieurs caractéristiques dans un champ <221> et un champ <222> annexe, et que ces caractéristiques s'accompagnent d'une numérotation négative ou positive, par exemple "PROPEP" ou "CHAIN", il convient d'employer la clé de caractérisation adéquate dans le listage des séquences selon la norme ST.26, par exemple "PROPEP" ou "CHAIN". On peut employer un qualificateur "note" pour reprendre les informations d'un champ <223>, le cas échéant;

b) Si le listage des séquences selon la norme ST.25 ne comporte pas de caractéristique dans un champ <221> et un champ <222> annexe, mais que la description de la demande contient des informations relatives à une numérotation négative ou positive, il convient d'employer la clé de caractérisation adéquate dans le listage des séquences selon la norme ST.26, par exemple "PROPEP" ou "CHAIN". Une autre possibilité consiste à utiliser la clé de caractérisation "REGION". On peut employer un qualificateur "note" pour reprendre les informations figurant dans la description de la demande, le cas échéant;

c) Si ni le listage des séquences selon la norme ST.25 ni la description de la demande ne contiennent d'informations sur une numérotation négative ou positive, pour éviter une éventuelle suppression d'élément dans le listage des séquences selon la norme ST.26, il convient d'utiliser la clé de caractérisation "REGION" lorsque l'emplacement de la caractéristique recouvre la région portant un numéro négatif dans la séquence conforme à la norme ST.25. Il faut aussi utiliser un qualificateur "note" pour indiquer que la séquence d'acides aminés portait un numéro négatif dans le listage des séquences établi selon la norme ST.25 aux fins de la demande dont la priorité est revendiquée.

Scénario 17

La norme ST.25 permet de fournir des informations concernant la publication dans les champs <300> à <313>. La norme ST.26 ne permet pas de communiquer ces informations.

Recommandation

Les informations figurant dans les champs <300> à <313> de la norme ST.25 doivent être reprises dans le corps de la demande annexe si elles n'y figurent pas déjà.

Scénario 18

La norme ST.25 ne prévoit pas de méthode normalisée pour indiquer qu'une région CDS d'une séquence de nucléotides doit être traduite au moyen d'un tableau de codes génétiques différent du tableau classique. Inversement, la norme ST.26 dispose d'un qualificateur "transl_table" qui peut être employé dans la clé de caractérisation "CDS" pour indiquer que la région doit être traduite au moyen d'un tableau de codes génétiques différent. Si le qualificateur "transl_table" n'est pas employé, on présume qu'il faut utiliser le tableau de codes génétiques classique.

Recommandations

a) Si le listage des séquences selon la norme ST.25 ou la description de la demande indique clairement qu'il faut traduire une région CDS au moyen d'un tableau de codes génétiques différent, il faut employer le qualificateur "transl_table" en précisant le numéro du tableau dans sa valeur. Ne pas employer ce qualificateur reviendrait à ajouter un élément car on présumerait qu'il faut utiliser le tableau classique ("Standard Code"). En outre, si l'on ne reprend pas les informations sur l'emploi d'un tableau différent dans le listage des séquences conforme à la norme ST.26, ces informations qui figuraient dans le listage des séquences selon la norme ST.25 ou dans la description de la demande vont probablement constituer des éléments supprimés.

b) Si le listage des séquences selon la norme ST.25 ou la description de la demande n'indique pas qu'il faut traduire une région CDS au moyen d'un tableau de codes génétiques différent, le qualificateur "transl_table" ne doit pas être utilisé; il peut toutefois l'être si la valeur du qualificateur est "1", ce qui désigne le tableau de codes classique. Toute indication d'une autre valeur dans ce qualificateur constituerait probablement un ajout ou une suppression d'élément.

Scénario 19

La norme ST.25 ne prévoit pas de méthode normalisée pour indiquer l'emplacement d'une caractéristique, notamment si celle-ci se trouve sur un site ou dans une région qui s'étend au-delà du résidu ou de la série de résidus définis, par exemple une région CDS d'une séquence de nucléotides qui s'étend au-delà de l'une ou des deux extrémités d'une séquence divulguée. En revanche, la norme ST.26 dispose d'un descripteur d'emplacement de la caractéristique qui offre une méthode normalisée pour indiquer l'emplacement d'un tel site ou d'une telle région en s'appuyant sur les symboles "<" ou ">". Ainsi, l'emplacement de la caractéristique "CDS" doit comporter le codon d'arrêt, même si celui-ci ne figure pas dans la séquence divulguée elle-même; à cette fin, on peut par exemple indiquer l'emplacement de la manière suivante : 1..>321.

Recommandations

a) Lorsque le listage des séquences selon la norme ST.25 n'indique pas explicitement que l'emplacement d'une caractéristique s'étend au-delà de la séquence, mais que cet emplacement ressort clairement de la divulgation ou de la séquence elle-même, par exemple si le codon d'arrêt d'une caractéristique CDS ne figure pas dans la séquence, on peut employer les symboles "<" ou ">" dans le listage des séquences selon la norme ST.26 sans ajouter d'élément.

b) Lorsque le listage des séquences selon la norme ST.25 n'indique pas explicitement que l'emplacement d'une caractéristique s'étend au-delà de la séquence, et que cet emplacement ne ressort pas clairement de la divulgation ou de la séquence elle-même, il peut être impossible de se conformer à la norme ST.26 sans ajouter d'élément. Dans ce cas, on peut faire valoir que la demande établissant une priorité et le listage des séquences sont eux-mêmes incomplets. La description de la caractéristique dans le listage des séquences selon la norme ST.26 n'obtiendra alors pas la priorité par rapport à la demande antérieure. Il faut veiller à fournir des informations complètes sur les caractéristiques lorsqu'on établit le listage des séquences et la divulgation de la demande originaux (selon la norme ST.25).

Scénario 20

L'appendice 1 de la norme ST.25 prévoit que si une séquence de nucléotides contient à la fois des fragments d'ADN et d'ARN, le champ <212> doit prendre la valeur "DNA" et la molécule combinée d'ADN/ARN doit être décrite plus en détail dans les champs <220> à <223>, qui sont consacrés aux caractéristiques. Cependant, la nature exacte de la description détaillée n'étant pas claire, cette obligation n'est généralement pas respectée. Aux termes du paragraphe 55 de la norme ST.26, 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.

Recommandations

a) Si le listage des séquences selon la norme ST.25 décrit des fragments d'ADN et d'ARN au moyen d'une ou

plusieurs caractéristiques en utilisant les clés de caractérisation <221> (misc_feature), <222> (Emplacement) et <223> (Autres informations) pour préciser si chaque fragment est de l'ADN ou de l'ARN, puis que ces informations sont converties au format ST.26 en utilisant une clé "misc_feature" pour chaque fragment d'ADN ou d'ARN, cette méthode ne devrait pas entraîner d'ajout d'élément.

b) Si le listage des séquences selon la norme ST.25 décrit des fragments d'ADN et d'ARN au moyen d'une ou plusieurs caractéristiques en utilisant une clé de caractérisation <221> autre que "misc_feature", ainsi que les clés <222> (Emplacement) et <223> (Autres informations) pour préciser si chaque fragment est de l'ADN ou de l'ARN, puis que ces informations sont converties au format ST.26 en utilisant une clé "misc_feature" pour chaque fragment d'ADN ou d'ARN et un qualificateur "note" supplémentaire avec la clé de caractérisation <221> d'origine comme valeur, cette méthode ne devrait pas entraîner d'ajout ni de suppression d'élément.

c) Si le listage des séquences selon la norme ST.25 indique l'identité (ADN ou ARN) et l'emplacement de chaque segment dans un champ <223> qui n'est pas lié à des champs <221> et <222>, par exemple pour préciser qu'une séquence est artificielle, la conversion de ces informations au format ST.26 utilisant une clé "misc_feature" pour chaque fragment d'ADN ou d'ARN ne devrait pas entraîner d'ajout d'élément.

d) Si le listage des séquences selon la norme ST.25 décrit la molécule dans une caractéristique au moyen d'une clé <221> (misc_feature) et d'une clé <223>, en précisant que la molécule combine de l'ADN et de l'ARN, mais qu'il ne fournit pas d'informations sur l'emplacement de chaque fragment, et

i) Si la description indique l'emplacement de chaque fragment d'ADN ou d'ARN, puis ces informations sont converties au format ST.26 en utilisant une clé "misc_feature" pour chaque fragment d'ADN ou d'ARN, cette méthode ne devrait pas entraîner d'ajout d'élément.

ii) Si la description n'indique pas l'emplacement de chaque fragment d'ADN ou d'ARN, il peut être impossible de se conformer à la norme ST.26 sans ajouter d'élément. Dans ce cas, on peut faire valoir que la demande établissant une priorité et le listage des séquences sont eux-mêmes incomplets. Les descriptions d'emplacement des caractéristiques dans le listage des séquences selon la norme ST.26 n'obtiendront alors pas la priorité par rapport à la demande antérieure. Il faut veiller à fournir des informations complètes sur les caractéristiques lorsqu'on établit le listage des séquences et la divulgation de la demande originaux (selon la norme ST.25).

e) Si le listage des séquences selon la norme ST.25 décrit la molécule dans une caractéristique au moyen d'une clé <221> autre que "misc_feature" et d'une clé <223>, en précisant que la molécule combine de l'ADN et de l'ARN, mais qu'il ne fournit pas d'informations sur l'emplacement de chaque fragment, et

i) Si la description indique l'emplacement de chaque fragment d'ADN ou d'ARN, puis ces informations sont converties au format ST.26 en utilisant une clé "misc_feature" pour chaque fragment d'ADN ou d'ARN et un qualificateur "note" supplémentaire prenant la valeur de la clé de caractérisation <221> originale, cette méthode ne devrait pas entraîner d'ajout d'élément.

ii) Si la description n'indique pas l'emplacement de chaque fragment d'ADN ou d'ARN, il peut être impossible de se conformer à la norme ST.26 sans ajouter d'élément. Dans ce cas, on peut faire valoir que la demande établissant une priorité et le listage des séquences sont eux-mêmes incomplets. Les descriptions d'emplacement des caractéristiques dans le listage des séquences selon la norme ST.26 n'obtiendront alors pas la priorité par rapport à la demande antérieure. Il faut veiller à fournir des informations complètes sur les caractéristiques lorsqu'on établit le listage des séquences et la divulgation de la demande originaux (selon la norme ST.25).

f) Si le listage des séquences selon la norme ST.25 indique dans un champ <223> (par exemple pour signaler une séquence artificielle) que la molécule est une combinaison d'ADN et d'ARN, mais qu'il ne fournit pas de clé de caractérisation ou d'informations sur l'emplacement de chaque fragment, et

i) Si la description indique l'emplacement de chaque fragment d'ADN ou d'ARN, puis ces informations sont converties au format ST.26 en utilisant une clé "misc_feature" pour chaque fragment d'ADN ou d'ARN, cette méthode ne devrait pas entraîner d'ajout d'élément.

ii) Si la description n'indique pas l'emplacement de chaque fragment d'ADN ou d'ARN, il peut être impossible de se conformer à la norme ST.26 sans ajouter d'élément. Dans ce cas, on peut faire valoir que la demande établissant une priorité et le listage des séquences sont eux-mêmes incomplets. Les descriptions d'emplacement des caractéristiques dans le listage des séquences selon la norme ST.26 n'obtiendront alors pas la priorité par rapport à la demande antérieure. Il faut veiller à fournir des informations complètes sur les caractéristiques lorsqu'on établit le listage des séquences et la divulgation de la demande originaux (selon la norme ST.25).

[Fin de l'annexe VII et de la norme]