F

OMPI
ORGANISATION MONDIALE
DE LA PROPRIÉTÉ
INTELLECTUELLE

CWS/4/7
ORIGINAL : ANGLAIS
DATE : 9 AVRIL 2014

Comité des normes de l'OMPI (CWS)

Quatrième session
Genève, 12 – 16 mai 2014

NOUVELLE NORME DE L'OMPI RELATIVE A LA PRÉSENTATION DES LISTAGES DES SÉQUENCES DE NUCLÉOTIDES ET D'ACIDES AMINÉS EN LANGAGE XML (*EXTENSIBLE MARKUP LANGUAGE*)

Document établi par le Secrétariat

1. À sa première session tenue en octobre 2010, le Comité des normes de l'OMPI (CWS) est convenu de créer la tâche n° 44, qui vise à établir une recommandation concernant la représentation des listages des séquences de nucléotides et d'acides aminés en langage XML (*eXtensible Markup Language*) pour adoption en tant que norme de l'OMPI. Le CWS a également décidé de créer une équipe d'experts chargée de mener à bien cette tâche (équipe d'experts SEQL). L'Office européen des brevets (OEB) a été désigné comme responsable de cette tâche. (Voir les paragraphes 27 à 30 du document CWS/1/10 et la tâche n° 44 dans le document CWS/3/12. L'annexe I du présent document contient également une description de cette tâche.)

2. Comme suite à cette décision du CWS, les représentants de 13 offices de propriété industrielle et le Bureau international ont été désignés pour participer à l'équipe d'experts. À ses deuxième et troisième sessions, le CWS a pris note du fait que l'OEB, en sa qualité de responsable de l'équipe d'experts, a rendu compte de l'état d'avancement des débats menés par cette équipe, et en particulier du programme de travail concernant l'élaboration des recommandations (voir les documents CWS/2/5 et CWS/3/6).

3. Après la troisième session du CWS, l'équipe d'experts SEQL a poursuivi ses débats par le biais du forum Wiki. Le rapport établi par le responsable de l'équipe d'experts sur les travaux menés par celle-ci figure à l'annexe I du présent document.

4. Comme suite à la demande susmentionnée du CWS, l'équipe d'experts SEQL a établi une proposition de nouvelle norme pour examen et approbation par le CWS. La proposition de titre de cette nouvelle norme est la suivante : "Norme ST.26 de l'OMPI – Recommandation de

norme concernant la représentation des listages des séquences de nucléotides et d'acides aminés en langage XML (*eXtensible Markup Language*)". Le projet de nouvelle norme ST.26 de l'OMPI, qui contient le corps du texte et cinq annexes, figure à l'annexe II du présent document.

5. L'équipe d'experts SEQL a également été chargée par le CWS de coordonner ses travaux avec l'organe compétent du PCT en ce qui concerne l'éventuelle incidence de la nouvelle norme ST.26 sur l'annexe C des Instructions administratives du PCT (voir le paragraphe 29 c) du document CWS/1/10). Les dispositions relatives au passage de la norme ST.25 à la nouvelle norme ST.26 de l'OMPI sont actuellement étudiées par les membres de l'équipe d'experts. Elles devraient être soumises au CWS pour examen à sa prochaine session, prévue en 2015 (voir le paragraphe 10, intitulé "Feuille de route", de l'annexe I du présent document).

6. Les offices de propriété intellectuelle sont priés de reporter les préparatifs en vue de la mise en œuvre de la nouvelle norme ST.26 de l'OMPI jusqu'à ce que ces dispositions sur le passage d'une norme à l'autre soient approuvées par le CWS. Dans l'intervalle, il faut continuer à appliquer la norme ST.25. Dès lors, et sous réserve que la nouvelle norme soit adoptée à la session en cours (quatrième session) du CWS, l'équipe d'experts propose de joindre la note de la rédaction suivante à la nouvelle norme :

"Note du Bureau international

"Le CWS est convenu de prier les offices de propriété industrielle de reporter les préparatifs en vue de la mise en œuvre de cette nouvelle norme ST.26 de l'OMPI jusqu'à ce que les recommandations relatives au passage de la norme ST.25 à la nouvelle norme ST.26 soient approuvées par le CWS à sa cinquième session, qui aura lieu en 2015. Dans l'intervalle, la norme ST.25 doit continuer d'être appliquée.

"La norme est publiée à des fins d'information des offices de propriété industrielle et d'autres parties intéressées.

"Le Comité des normes de l'OMPI (CWS) a adopté la présente norme à [sa quatrième session tenue le 16 mai 2014]."

7. Le CWS est invité à :

a) prendre note du rapport sur l'état d'avancement des travaux effectués par l'équipe d'experts SEQL qui fait l'objet de l'annexe I du présent document;

b) adopter la "norme ST.26 de l'OMPI – Recommandation de norme concernant la représentation des listages des séquences de nucléotides et d'acides aminés en langage XML (eXtensible Markup Language)" comme titre de la proposition de norme;

c) examiner et adopter la norme ST.26 de l'OMPI, qui fait l'objet de l'annexe II du présent document;

d) examiner et approuver la note de la rédaction qui doit être jointe à la norme ST.26 de l'OMPI (voir le paragraphe 6 ci-dessus); et

e) prier l'équipe d'experts SEQL d'établir une proposition de dispositions concernant le passage d'une norme à l'autre, conformément au paragraphe 5 ci-dessus, et de soumettre ces dispositions au CWS pour examen et approbation à sa cinquième session.

[Les annexes suivent]

RAPPORT SUR L'ELABORATION D'UNE NOUVELLE NORME DE L'OMPI RELATIVE A LA PRESENTATION DES LISTAGES DES SEQUENCES DE NUCLEOTIDES ET D'ACIDES AMINES EN LANGAGE XML (*EXTENSIBLE MARKUP LANGUAGE*)

Document établi par l'Office européen des brevets (OEB)

RAPPEL

1. L'Équipe d'experts chargée de la norme relative aux listages des séquences a été créée par le Comité des normes de l'OMPI (CWS) à sa première session (tenue du 25 au 29 octobre 2010) afin de traiter de la tâche n° 44 (voir le paragraphe 29 du document CWS/1/10) :

> "Établir une recommandation concernant la représentation des listages des séquences de nucléotides et d'acides aminés en langage XML (*eXtensible Markup Language*) pour adoption en tant que norme de l'OMPI. La proposition relative à l'établissement de cette nouvelle norme de l'OMPI devrait être assortie d'une étude de l'incidence de ladite norme sur la norme ST.25 actuelle de l'OMPI, indiquant notamment les modifications à apporter à la norme ST.25."

2. L'équipe d'experts a également été priée :

> "de coordonner ses travaux avec l'organe compétent du PCT en ce qui concerne l'incidence éventuelle de ladite norme sur l'annexe C des Instructions administratives du PCT."

3. Le rôle de responsable de l'équipe d'experts a été attribué à l'Office européen des brevets (OEB), qui a depuis lors tenu six séries de discussions sur le Wiki de l'OMPI et a présenté une version finale de la norme à des fins de consultation publique. Le principe de différenciation des aspects techniques de la norme ST.25 de ceux de l'annexe C (Instructions administratives du PCT) a fait l'objet d'un accord à la dix-huitième session de la réunion des administrations internationales en février 2011 (voir les paragraphes 88 à 92 du document PCT/MIA/18/16) et à la quatrième session du Groupe de travail du PCT en juin 2011 (voir les paragraphes 180 à 188 du document PCT/WG/4/17).

4. Compte tenu des observations communiquées par les membres de l'équipe d'experts, une série finale de discussions a eu lieu pour trouver un accord sur les exigences de la norme.

RAPPORT SUR L'ETAT D'AVANCEMENT DES TRAVAUX

5. L'équipe d'experts a commencé à fonctionner en février 2011 sur la base des projets élaborés par l'OEB. De nombreux offices ont participé au processus et publié des observations utiles sur le Wiki correspondant de l'OMPI.

6. En mars 2012, l'équipe d'experts a achevé la mise au point d'une version préliminaire de la norme, qui a pu être utilisée par les offices pour consulter leurs utilisateurs respectifs. Plusieurs questions essentielles ont été soulevées dans les commentaires des utilisateurs et ont été résolues en coopération avec les fournisseurs des bases de données DDBJ, EBI et NCBI.

7. La sixième série de discussions s'est achevée en septembre 2013 et la version de la norme reprenant les améliorations issues des consultations publiques et des débats ultérieurs entre les membres de l'équipe d'experts et les fournisseurs de bases de données a été publiée sur le Wiki de l'OMPI pour examen final.

8. Compte tenu des observations communiquées par les membres de l'équipe d'experts, une série finale de discussions a été tenue pour trouver un accord sur les exigences de la norme. L'équipe d'experts a provisoirement intitulé cette norme ST.26. Le corps du texte et les annexes soumis par l'équipe d'experts au CWS pour examen et approbation contiennent les améliorations suivantes par rapport à la norme ST.25 actuelle :

a) toutes les questions procédurales du PCT sont intégrées dans les instructions administratives du PCT : la nouvelle norme se concentre sur les aspects techniques pour permettre une présentation optimale des listages des séquences (la partie relative à la biotechnologie) et pour adopter le format approprié de la présentation (en l'occurrence le XML);

b) la partie relative à la biotechnologie a été considérablement améliorée afin de tenir compte des normes de l'industrie moderne, par exemple :

- l'inclusion des nucléotides et des acides aminés modifiés n'ayant pas été précédemment prévus (par exemple, les D-aminoacides, les acides nucléiques peptidiques, les morpholinos, etc.) qui ont pris de l'importance dans l'industrie et doivent pouvoir se prêter à une recherche électronique;

- des instructions claires en ce qui concerne les séquences pourvues de brèches et les variantes de séquence;

- une clarification concernant les caractéristiques et les annotations;

- une cohérence avec les dernières exigences en matière de consortiums de bases de données publiques de séquences biologiques (INSDC et UniProt); et

- la définition XML est spécifique et indépendante des normes ST.36 ou ST.96.

c) La syntaxe prévue dans la définition de type de document (DTD) employée dans la norme ST.26 améliore la précision des données et permet d'effectuer un contrôle automatique de la qualité des données.

9. L'équipe d'experts poursuivra ses travaux sur les questions liées au passage de la norme ST.25 à la norme ST.26 en 2014 et 2015 afin de soumettre ses recommandations à cet égard au CWS pour examen et approbation à la cinquième session de celui-ci.

FEUILLE DE ROUTE

10. Une nouvelle série de discussions sera menée après la quatrième session du CWS et portera sur l'établissement des recommandations concernant le passage d'une norme à l'autre. Ces recommandations devront être présentées au CWS à sa session de 2015.

[L'annexe II suit]

NORME ST.26

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

Version finale
Proposition présentée par l'équipe d'experts SEQL pour examen et adoption par le CWS à sa quatrième session

TABLE DES MATIÈRES

ANNEXES

NORME ST.26

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

Version finale
Proposition présentée par l'équipe d'experts SEQL pour examen et adoption par le CWS à sa quatrième session

INTRODUCTION

1. La présente norme définit la manière dont des séquences de nucléotides et d'acides aminés doivent être divulguées dans une demande de brevet pour pouvoir être jointes à un listage des séquences. Elle précise les caractéristiques de ces divulgations et la définition de type de document (DTD) à employer lorsque le listage des séquences est effectué au format XML (*eXtensible Markup Language*). Il est recommandé que les offices de propriété industrielle acceptent tous les listages de séquences conformes à cette norme qui sont déposés en tant que partie intégrante d'une demande de brevet ou en relation avec une demande de brevet.

2. Cette norme a pour but :

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 D-aminoacides et les acides aminés contenant des chaînes latérales modifiées ou synthétiques. Les acides aminés seront considérés comme des L-aminoacides non modifiés sauf s'il est précisé dans leur description qu'ils sont modifiés au sens du paragraphe 29;

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) "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;

d) "nucléotide" désigne tout nucléotide ou analogue nucléotidique qui peut être représenté à l'aide des symboles indiqués dans l'annexe I (voir section 1, tableau 1). Les nucléotides peuvent notamment contenir une base pyrimidique ou purine modifiée ou synthétique, ou un ribose ou désoxyribose modifié ou synthétique, et peuvent être reliés par une liaison inter-nucléoside de 3' à 5' modifiée ou synthétique, c'est-à-dire par toute fraction chimique assurant la même fonction structurelle que la fraction phosphate de l'ADN ou de l'ARN, par exemple la fraction phosphorothioate;

e) "résidu" désigne tout nucléotide ou acide aminé individuel dans une séquence;

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

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

h) "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;

i) "inconnu", pour un nucléotide ou un acide aminé, signifie qu'un seul nucléotide ou acide aminé est présent mais que son identité est inconnue ou non divulguée.

PORTÉE

4. 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.

5. 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.

6. Aux fins de la présente norme, une séquence doit être intégrée dans un listage si elle est divulguée dans n'importe quelle partie d'une demande de brevet par l'énumération de ses résidus, et s'il s'agit :

a) d'une séquence non ramifiée ou d'une partie linéaire d'une séquence ramifiée contenant au moins 10 nucléotides définis de manière spécifique, et dont les nucléotides adjacents ont des liaisons de 3' à 5' (ou de 5' à 3'), ou

b) d'une séquence non ramifiée ou d'une partie linéaire d'une séquence ramifiée contenant au moins quatre acides aminés définis de manière spécifique, et dont les acides aminés adjacents ont des liaisons peptidiques.

7. Un listage des séquences ne doit contenir aucune séquence comportant moins de 10 nucléotides définis de manière spécifique ou moins de quatre acides aminés définis de manière spécifique.

RÉFÉRENCES

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

Collaboration internationale sur les bases de données de séquences de nucléotides (INSDC)	http://www.insdc.org/;
Norme ISO 639−1 – Codes pour la représentation des noms de langue	Partie 1 : Code Alpha2;
Consortium UniProt	http://www.uniprot.org/;
Norme du W3C sur le XML 1.0	http://www.w3.org/;
Norme ST.2 de l'OMPI	Indication normalisée des dates à l'aide du calendrier grégorien;
Norme ST.3 de l'OMPI	Codes à deux lettres pour la représentation des États, autres entités et organisations intergouvernementales.
Norme ST.16 de l'OMPI	Identification de différents types de documents de brevet;
Norme ST.25 de l'OMPI	Présentation des listages des séquences de nucléotides et d'acides aminés.

PRÉSENTATION DES SÉQUENCES

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

Séquences de nucléotides

10. Toute séquence de nucléotides doit être représentée par un seul brin de codage, dans le sens 5'-3' et de gauche à droite. Les désignations 5' et 3' ne doivent pas apparaître dans la séquence. Toute séquence de nucléotides représentée par deux brins de codage et divulguée par énumération des résidus des deux brins doit être présentée sous la forme :

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.

11. La numérotation des positions des nucléotides doit commencer par la première base de la séquence, qui portera le numéro 1. Elle doit être continue dans toute la séquence dans le sens 5'-3'.

12. Cette méthode de numérotation des séquences de nucléotides s'applique aussi aux séquences de nucléotides de configuration circulaire. Dans ce cas, le déposant doit choisir le nucléotide correspondant au premier numéro.

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 au sens du paragraphe 18.

15. Lorsqu'il convient d'employer un symbole ambigu (représentant deux bases possibles ou plus), il faut choisir le symbole le plus restrictif. Si par exemple une base dans une position quelconque pouvait être désignée par "a" ou "g", il faut employer "r" au lieu de "n". Le symbole "n" sera considéré comme équivalent à l'un des symboles "a", "c", "g" ou "t/u", sauf s'il est accompagné d'une description supplémentaire au sens des paragraphes 16 et 17 ou 20. Ce symbole "n" ne peut être employé que pour représenter un nucléotide. Il peut représenter un seul nucléotide modifié ou "inconnu" s'il est accompagné d'une description supplémentaire au sens des paragraphes 16 et 17 ou 20.

16. Les nucléotides modifiés doivent être représentés dans la séquence comme les bases non modifiées correspondantes, c'est-à-dire par "a", "c", "g" ou "t" chaque fois que possible. Tout nucléotide modifié apparaissant dans une séquence et ne pouvant être représenté à l'aide d'un autre symbole indiqué dans l'annexe I (voir section 1, tableau 1), comme par exemple un nucléotide n'existant pas à l'état naturel, doit être représenté par le symbole "n". Lorsque ce symbole "n" est employé pour représenter un nucléotide modifié, il n'est l'équivalent que d'un seul résidu.

17. Tout nucléotide modifié doit être accompagné d'une description supplémentaire dans le tableau de caractéristiques (voir les paragraphes 59 et suivants) comportant la clé de caractérisation "modified_base" et le qualificateur obligatoire "mod_base". La valeur qualificative ne peut être constituée que d'une seule abréviation issue de l'annexe I (voir section 2, tableau 2). Si cette abréviation est "OTHER", le nom complet non abrégé de la base modifiée doit être indiqué dans un qualificateur de type "note". Les abréviations (ou les noms complets) indiquées dans l'annexe I (voir section 2, tableau 2) qui sont mentionnées ci-dessus ne doivent pas être employées dans la séquence elle-même.

18. 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".

19. Les exemples ci-après illustrent la manière dont des nucléotides modifiés doivent être présentés pour être conformes aux paragraphes 16 et 17 ci-dessus :

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

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

Exemple 2 : Nucléotide modifié représenté par la valeur "OTHER" indiquée dans l'annexe I (voir section 2, tableau 2)

```
<INSDFeature>
   <INSDFeature_key>modified_base</INSDFeature_key>
   <INSDFeature_location>4</INSDFeature_location>
   <INSDFeature_quals>
       <INSDQualifier>
           <INSDQualifier_name>mod_base</INSDQualifier_name>
           <INSDQualifier_value>OTHER</INSDQualifier_value>
       </INSDQualifier>
       <INSDQualifier>
           <INSDQualifier_name>note</INSDQualifier_name>
           <INSDQualifier_value>xanthine</INSDQualifier_value>
       </INSDQualifier>
   </INSDFeature_quals>
</INSDFeature>
```

20. 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.

21. Toute région contenant un nombre connu de résidus “a”, “c”, “g”, “t” ou “n” auxquels la même description s’applique peut faire l’objet d’une description commune au moyen de la syntaxe “x..y” dans le descripteur d’emplacement de l’élément `INSDFeature_location` (voir les paragraphes 65 à 72). On trouvera des détails sur la présentation des variantes de séquence, par exemple des suppressions, des adjonctions ou des remplacements, aux paragraphes 92 à 97.

22. L’exemple ci-après illustre la présentation d’une région de nucléotides modifiés faisant l’objet de la même description au sens du paragraphe 21 ci-dessus :

```
<INSDFeature>
   <INSDFeature_key>modified_base</INSDFeature_key>
   <INSDFeature_location>358..485</INSDFeature_location>
   <INSDFeature_quals>
           <INSDQualifier>
           <INSDQualifier_name>mod_base</INSDQualifier_name>
           <INSDQualifier_value>OTHER</INSDQualifier_value>
       </INSDQualifier>
           <INSDQualifier>
           <INSDQualifier_name>note</INSDQualifier_name>
           <INSDQualifier_value>isoguanine</INSDQualifier_value>
       </INSDQualifier>
   </INSDFeature_quals>
</INSDFeature>
```

Séquences d’acides aminés

23. Les acides aminés d’une séquence protéinique ou peptidique doivent être énumérés dans le sens amino-carboxy et de gauche à droite. Les groupes amino et carboxy ne doivent pas être représentés dans la séquence.

24. La numérotation des positions des acides aminés doit commencer au premier acide aminé de la séquence, numéroté 1, en tenant compte des acides aminés précédant la protéine mature, par exemple les préséquences, les proséquences et les pré-proséquences ainsi que les séquences signal. Elle doit être continue sur l’ensemble de la séquence dans le sens amino-carboxy.

25. 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.

26. Lorsqu’il convient d’employer un symbole ambigu (représentant deux acides aminés possibles ou plus), il faut choisir le symbole le plus restrictif. Si par exemple un acide aminé dans une position quelconque pouvait être de l’acide aspartique ou de l’asparagine, il faut employer le symbole “B” au lieu de “X”. Le symbole “X” ne sera pas considéré comme équivalent à l’un des symboles “A”, “R”, “N”, “D”, “C”, “Q”, “E”, “G”, “H”, “I”, “L”, “K”, “M”, “F”, “P”, “O”, “S”, “U”, “T”, “W”, “Y” ou “V”, sauf s’il est accompagné d’une description supplémentaire au sens des paragraphes 28 à 30 ou 31 à 33. Le symbole “X” ne peut être employé que pour représenter un acide aminé. Il peut représenter un seul acide aminé s’il est accompagné d’une description supplémentaire au sens des paragraphes 28 à 30 ou 31 à 33. On trouvera des détails sur la présentation des variantes de séquence, par exemple des suppressions, des adjonctions ou des remplacements, aux paragraphes 92 à 97.

27. Les séquences d’acides aminés séparées par un ou plusieurs espaces blancs ou symboles internes de fin (par exemple “Ter”, l’astérisque ou le point) doivent être présentées comme des séquences distinctes pour chaque séquence qui contient au moins quatre acides aminés définis de manière spécifique et qui est visée par le paragraphe 6. Chacune de ces séquences distinctes doit être présentée, dans le listage des séquences, avec son propre numéro d’identification de séquence et uniquement à l’aide des symboles indiqués dans l’annexe I (voir section 3, tableau 3). Les symboles de fin et les espaces blancs ne doivent pas être employés dans les séquences figurant dans un listage.

28. Les acides aminés modifiés, y compris les D-aminoacides, doivent être représentés dans la séquence comme les acides aminés non modifiés correspondants chaque fois que possible. Tout acide aminé modifié apparaissant dans une séquence et ne pouvant être représenté à l’aide d’un autre symbole indiqué dans l’annexe I (voir section 3, tableau 3), doit être représenté par le symbole “X”. Ce symbole “X” n’est l’équivalent que d’un seul résidu.

29. Tout acide aminé modifié doit être accompagné d’une description supplémentaire dans le tableau de caractéristiques (voir les paragraphes 60 et suivants). Il convient d’employer la clé de caractérisation “MOD_RES” et le qualificateur “NOTE” pour les acides aminés modifiés après traduction, et la clé de caractérisation “SITE” ainsi que le qualificateur “NOTE” pour les autres acides aminés modifiés. La valeur du qualificateur “NOTE” doit être soit une abréviation indiquée dans l’annexe I (voir section 4, tableau 4), soit le nom complet non abrégé de l’acide aminé modifié. Les abréviations indiquées dans le tableau 4 précité ou les noms complets non abrégés ne doivent pas être employés dans la séquence elle-même.

30. Les exemples ci-après illustrent la manière dont des acides aminés modifiés doivent être présentés pour être conformes au paragraphe 29 ci-dessus :

Exemple 1 : Acide aminé modifié après traduction

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

Exemple 2 : Acide aminé modifié différemment

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

Exemple 3 : D-aminoacide

```
<INSDFeature>
   <INSDFeature_key>SITE</INSDFeature_key>
   <INSDFeature_location>9</INSDFeature_location>
   <INSDFeature_quals>
       <INSDQualifier>
           <INSDQualifier_name>NOTE</INSDQualifier_name>
           <INSDQualifier_value>D-Arginine</INSDQualifier_value>
       </INSDQualifier>
   </INSDFeature_quals>
</INSDFeature>
```

31. Tout acide aminé “inconnu” ou “autre” qui n’est pas visé par le paragraphe 28 doit être représenté par le symbole “X” dans la séquence. Le symbole “X” ne peut être l’équivalent que d’un seul résidu.

32. Tout acide aminé “inconnu” désigné par “X” doit être accompagné d’une description supplémentaire dans le tableau de caractéristiques (voir paragraphes 60 et suivants) comportant la clé de caractérisation “UNSURE” et éventuellement le qualificateur “NOTE”. Tout acide aminé “autre” désigné par “X” doit être accompagné d’une description supplémentaire comportant la clé de caractérisation “SITE” ou “MOD_RES”, selon le cas, ainsi que le qualificateur “NOTE” avec le nom complet non abrégé de l’acide aminé “autre”.

33. Les exemples ci-après illustrent la manière dont des acides aminés “inconnus” ou “autres” doivent être présentés pour être conformes aux paragraphes 31 et 32 ci-dessus :

Exemple 1 : Acide aminé “inconnu”

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

Exemple 2 : Acide aminé “autre”

```
<INSDFeature>
   <INSDFeature_key>SITE</INSDFeature_key>
   <INSDFeature_location>3</INSDFeature_location>
   <INSDFeature_quals>
```

```
            <INSDQualifier>
                <INSDQualifier_name>NOTE</INSDQualifier_name>
                <INSDQualifier_value>Homoserine</INSDQualifier_value>
            </INSDQualifier>
        </INSDFeature_quals>
    </INSDFeature>
```

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 65 à 71). On trouvera des détails sur la présentation des variantes de séquence, par exemple des suppressions, des adjonctions ou des remplacements, aux paragraphes 92 à 97.

Présentation de cas particuliers

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 comme une séquence unique avec un numéro d’identification de séquence unique.

36. Toute séquence divulguée par énumération de ses résidus qui contient des régions de résidus énumérés de manière spécifique et séparés par une ou plusieurs régions de résidus “n” ou “X” contigus (voir respectivement les paragraphes 15 et 26), et pour laquelle le nombre exact de résidus dans chaque région est divulgué, doit être ajoutée au listage des séquences comme une séquence unique avec un numéro d’identification de séquence unique.

37. Toute séquence divulguée par énumération de ses résidus qui contient des régions de résidus énumérés de manière spécifique et séparés par une ou plusieurs brèches composées d’un nombre inconnu ou non divulgué de résidus doit être insérée dans le listage des séquences comme une série de séquences distinctes. Chacune de ces séquences distinctes doit comporter une région de résidus énumérés de manière spécifique et disposer de son propre numéro d’identification de séquence. Le nombre de séquences distinctes doit ainsi être égal au nombre de régions de résidus énumérés de manière spécifique. Les séquences contenant des brèches composées d’un nombre inconnu ou non divulgué de résidus ne doivent pas être ajoutées au listage des séquences en tant que séquences uniques.

STRUCTURE DU LISTAGE DE SÉQUENCES EN XML

38. En application du paragraphe 5 ci-dessus, l’instance XML d’un fichier de listage des séquences conforme à la présente norme se compose des éléments suivants :

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 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 1.0//EN”
“ST26SequenceListing_V1_0.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 peuvent comporter n’importe quel caractère Unicode à l’exception des caractères réservés, qui doivent être remplacés selon la méthode décrite au paragraphe 41;

b) les informations figurant dans tous les autres éléments de la partie consacrée aux informations générales et dans tous les éléments de la partie sur les données des séquences

- doivent être composées de caractères imprimables (y compris le caractère d’espace) indiqués dans le tableau de codes des caractères latins de base de la norme Unicode, à l’exception des caractères réservés, qui doivent être remplacés selon la méthode décrite au paragraphe 41 (c’est-à-dire que ces caractères sont limités aux points de codes Unicode 0020, 0021, 0023 jusqu’à 0026, 0028 jusqu’à 003B, 003D, et 003F jusqu’à 007E – voir l’annexe IV), et

- les seules entités de caractères autorisées sont les entités prédéfinies prévues au paragraphe 41.

41. Dans l'instance XML d'un listage des séquences, les caractères réservés suivants doivent être remplacés par les entités prédéfinies correspondantes lorsqu'ils sont employés pour renseigner la valeur d'un attribut ou le contenu d'un élément :

Caractère réservé	Entités prédéfinies
<	&lt;
>	&gt;
&	&amp;
"	&quot;
'	&apos;

On trouvera un exemple au paragraphe 72.

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.

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_0".	Obligatoire
`fileName`	Nom du fichier contenant le listage des séquences.	Facultatif
`softwareName`	Nom du logiciel ayant créé le fichier.	Facultatif
`softwareVersion`	Version du logiciel ayant créé le fichier.	Facultatif
`productionDate`	Date de production du fichier contenant le listage des séquences (format "SSAA-MM-JJ").	Facultatif

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_0" fileName="US11_405455_SEQL.xml"
softwareName="SEQL-software-name" softwareVersion="1.0" productionDate="2006-05-10">
      {...}*
</ST26SequenceListing>

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

Partie consacrée aux informations générales

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	Identification de la demande fournie par l'office de dépôt (ex : PCT/IB2013/099999)	Obligatoire
FilingDate	Date de dépôt de la demande de brevet pour laquelle le listage des séquences est remis (au format ST.2 "SSAA-MM-JJ", en désignant l'année civile sur 4 chiffres, le mois civil sur 2 chiffres et le jour du mois civil sur 2 chiffres, p. ex. 2015-01-31)	Obligatoire lorsqu'un listage des séquences est remis à un moment quelconque après l'attribution d'une date de dépôt.
ApplicantFileReference	Identificateur unique attribué par le demandeur pour désigner une demande particulière, composé de caractères définis au paragraphe 40 b).	Obligatoire lorsqu'un listage des séquences est remis à un moment quelconque avant l'attribution du numéro de demande; facultatif dans les autres cas.
EarliestPriorityApplicationIdentification	Identification de la première revendication de priorité dans la demande (contient également les éléments IPOfficeCode, ApplicationNumberText et FilingDate, voir ApplicationIdentification ci-dessus)	Obligatoire si une priorité est revendiquée.
ApplicantName	Nom du premier déposant mentionné, composé de caractères définis au paragraphe 40 a). Cet élément comporte l'attribut obligatoire languageCode conformément au paragraphe 47.	Obligatoire
ApplicantNameLatin	Si l'élément ApplicantName comporte des caractères différents de ceux définis au paragraphe 40 b), une traduction ou une translittération du nom du premier déposant mentionné doit être fournie et doit aussi se composer de caractères définis au paragraphe 40 b).	Obligatoire si l'élément ApplicantName contient des caractères non latins.
InventorName	Nom du premier inventeur mentionné, composé de caractères définis au paragraphe 40 a). Cet élément comporte l'attribut obligatoire languageCode conformément au paragraphe 47.	Facultatif

Élément	Description	Obligatoire/ Facultatif
`InventorNameLatin`	Si l'élément `InventorName` comporte des caractères différents de ceux définis au paragraphe 40 b), une traduction ou une translittération du nom du premier inventeur mentionné doit être fournie et doit aussi se composer de caractères définis au paragraphe 40 b).	Facultatif
`InventionTitle`	Titre de l'invention, composé de caractères définis au paragraphe 40 a) dans la langue de dépôt. Une traduction du titre de l'invention dans d'autres langues peut être fournie; elle doit alors se composer de caractères définis au paragraphe 40 a) et apparaître sous des éléments `InventionTitle` supplémentaires. Cet élément comporte l'attribut obligatoire `languageCode` défini au paragraphe 48. Le titre de l'invention doit comporter de préférence deux à sept mots.	Obligatoire dans la langue de dépôt. Facultatif dans d'autres langues.
`SequenceTotalQuantity`	Nombre total de toutes les séquences apparaissant dans le listage, y compris les séquences délibérément omises (également appelées séquences vides) (voir le paragraphe 9).	Obligatoire

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 1.0//EN"
"ST26SequenceListing_V1_0.dtd">
<ST26SequenceListing dtdVersion="V1_0" fileName="Invention_SEQL.xml"
softwareName="SEQL-software-name" softwareVersion="1.0" productionDate="2015-05-10">
   <ApplicantFileReference>AB123</ApplicantFileReference>
   <EarliestPriorityApplicationIdentification>
       <IPOfficeCode>IB</IPOfficeCode>
       <ApplicationNumberText>PCT/IB2013/099999</ApplicationNumberText>
           <FilingDate>2014-07-10</FilingDate>
   </EarliestPriorityApplicationIdentification>
   <ApplicantName languageCode="EN">GENOS Co., Inc.</ApplicantName>
   <InventorName languageCode="EN">Keiko Nakamura</InventorName>
   <InventionTitle languageCode="EN">SIGNAL RECOGNITION PARTICLE RNA AND
PROTEINS</InventionTitle>
   <SequenceTotalQuantity>9</SequenceTotalQuantity>
   <SequenceData sequenceIDNumber="1"> {...}* </SequenceData>
   <SequenceData sequenceIDNumber="2"> {...} </SequenceData>
   <SequenceData sequenceIDNumber="3"> {...} </SequenceData>
   <SequenceData sequenceIDNumber="4"> {...} </SequenceData>
   <SequenceData sequenceIDNumber="5"> {...} </SequenceData>
   <SequenceData sequenceIDNumber="6"> {...} </SequenceData>
   <SequenceData sequenceIDNumber="7"> {...} </SequenceData>
   <SequenceData sequenceIDNumber="8"> {...} </SequenceData>
   <SequenceData sequenceIDNumber="9"> {...} </SequenceData>
</ST26SequenceListing>

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

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

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

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

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 le paragraphe 8 b)) doit être indiqué dans l'attribut `languageCode` pour chaque élément. Si le nom du déposant contient des caractères différents de l'alphabet latin tel que défini au paragraphe 40 b), une translittération ou une traduction de ce nom doit aussi être fournie en caractères latins dans l'élément `ApplicantNameLatin`. Si le nom de l'inventeur contient des caractères différents de l'alphabet latin, une translittération ou une traduction de ce nom doit aussi être fournie en caractères latins dans l'élément `InventorNameLatin`.

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 le paragraphe 8 b)) 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 languageCode ="JA">特許　太郎</InventorName>
<InventorNameLatin>Taro Tokkyo</InventorNameLatin>
<InventionTitle languageCode="JA"> efg タンパク質のためのマウス abcd-1 遺伝子
</InventionTitle>
<InventionTitle languageCode="EN"> Mus musculus abcd-1 gene for efg protein
</InventionTitle>
<InventionTitle languageCode="FR"> Gène abcd-1 de Mus musculus pour protéine efg
</InventionTitle>
```

Partie consacrée aux données sur les séquences

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 9), 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 présentée. Pour les séquences de nucléotides, le type de molécule doit être ADN ou ARN. Pour les séquences protéiniques ou peptidiques, le type de molécule doit être AA. (Cet élément est distinct des qualificateurs "mol_type" et "MOL_TYPE" mentionnés aux paragraphes 55 et 85.) Par exemple :

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

55. Si une séquence de nucléotides contient à la fois des fragments d'ADN et d'ARN, l'élément INSDSeq_moltype doit prendre la valeur "DNA". La molécule combinée d'ADN/ARN doit en outre être décrite dans le tableau de caractéristiques à l'aide de la clé de caractérisation "source", du qualificateur obligatoire "organism", qui prend la valeur "synthetic construct", et du qualificateur obligatoire "mol_type", qui prend la valeur "other DNA". Chaque fragment d'ADN et d'ARN de la molécule combinée d'ADN/ARN doit en outre être décrit par la clé de caractérisation "misc_feature" et par le qualificateur "note", ce dernier indiquant s'il s'agit d'un fragment d'ADN ou d'ARN.

56. L'exemple ci-après illustre la description d'une séquence de nucléotides contenant à la fois des fragments d'ADN et d'ARN comme le prévoit le paragraphe 55 ci-dessus :

```
<INSDSeq>
   <INSDSeq_length>120</INSDSeq_length>
   <INSDSeq_moltype>DNA</INSDSeq_moltype>
   <INSDSeq_division>PAT</INSDSeq_division>
   <INSDSeq_feature-table>
       <INSDFeature>
           <INSDFeature_key>source</INSDFeature_key>
           <INSDFeature_location>1..120</INSDFeature_location>
           <INSDFeature_quals>
               <INSDQualifier>
                   <INSDQualifier_name>organism</INSDQualifier_name>
                   <INSDQualifier_value>synthetic construct</INSDQualifier_value>
               </INSDQualifier>
               <INSDQualifier>
                   <INSDQualifier_name>mol_type</INSDQualifier_name>
                   <INSDQualifier_value>other DNA</INSDQualifier_value>
               </INSDQualifier>
           </INSDFeature_quals>
       </INSDFeature>
       <INSDFeature>
           <INSDFeature_key>misc_feature</INSDFeature_key>
```

```
                <INSDFeature_location>1..60</INSDFeature_location>
                <INSDFeature_quals>
                    <INSDQualifier>
                        <INSDQualifier_name>note</INSDQualifier_name>
                        <INSDQualifier_value>DNA fragment</INSDQualifier_value>
                    </INSDQualifier>
                </INSDFeature_quals>
            </INSDFeature>
            <INSDFeature>
                <INSDFeature_key>misc_feature</INSDFeature_key>
                <INSDFeature_location>61..120</INSDFeature_location>
                <INSDFeature_quals>
                    <INSDQualifier>
                        <INSDQualifier_name>note</INSDQualifier_name>
                        <INSDQualifier_value>RNA fragment</INSDQualifier_value>
                    </INSDQualifier>
                </INSDFeature_quals>
            </INSDFeature>
        </INSDSeq_feature-table>
        <INSDSeq_sequence>
          cgacccacgcgtccgaggaaccaaccatcacgtttgaggacttcgtgaaggaattggataatacccgtccctaccaaaat
    ggcgagcgccgactcattgctcctcgtaccgtcgagcggc
        </INSDSeq_sequence>
    </INSDSeq>
```

57. L'élément `INSDSeq_sequence` doit divulguer la séquence. Les résidus de la séquence doivent être présentés de manière contiguë à l'aide des symboles adéquats indiqués dans l'annexe I (voir section 1, tableau 1 et section 3, tableau 3). La séquence ne doit pas contenir de chiffres, de signes de ponctuation ou d'espaces blancs.

58. Une séquence délibérément omise doit être présentée 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 présentée pour être conforme au paragraphe 58 ci-dessus :

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

Tableau de caractéristiques

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 présentée correspondant à la caractéristique	Obligatoire
INSDFeature_quals	Qualificateur contenant des informations complémentaires sur une caractéristique	Obligatoire si la clé de caractérisation nécessite un ou plusieurs qualificateurs, p. ex. “source”. Facultatif dans les autres cas.

Clés de caractérisation

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 la clé de caractérisation “SOURCE” est obligatoire pour toutes les séquences d’acides aminés, sauf s’il s’agit d’une séquence délibérément omise. Chaque séquence doit comporter une clé “source” ou “SOURCE” unique couvrant la séquence tout entière. Si une séquence provient de plusieurs sources, celles-ci doivent en outre être décrites dans le tableau de caractéristiques à l’aide de la clé de caractérisation “misc_feature” et du qualificateur “note” pour les séquences de nucléotides, et de la clé de caractérisation “REGION” et du qualificateur “NOTE” pour les séquences d’acides aminés.

64. Certaines clés de caractérisation nécessitent l’emploi d’une clé de caractérisation supplémentaire, appelée “Parent Key”. Ainsi, la clé de caractérisation “C_region” doit être accompagnée de la clé “CDS” (voir annexe I, section 5).

Emplacement de la caractéristique

65. L’élément obligatoire `INSDFeature_location` doit contenir au moins un descripteur d’emplacement qui définit un site ou une région correspondant à une caractéristique de la séquence dans l’élément `INSDSeq_sequence`. Il peut contenir plusieurs descripteurs d’emplacement (voir les paragraphes 68 à 71).

66. Le descripteur d’emplacement peut être un numéro de résidu unique, un site entre deux numéros de résidus adjacents, une région délimitant une série de numéros de résidus contigus, ou un site ou une région qui s’étend au-delà du résidu ou de la série de résidus particuliers. On peut employer plusieurs descripteurs d’emplacement en conjonction avec un opérateur d’emplacement quand une caractéristique correspond à des sites ou des régions discontinus de la séquence (voir les paragraphes 68 à 71). Le descripteur d’emplacement ne doit pas comporter de numéros de résidus en dehors de la série indiquée pour la séquence dans l’élément `INSDSeq_sequence`.

67. La syntaxe de chaque type de descripteurs d’emplacement est indiquée dans le tableau ci-dessous, où x et y sont des numéros de résidus indiqués en nombres entiers non négatifs et inférieurs ou égaux à la longueur de la séquence dans l’élément `INSDSeq_sequence`, et où x est inférieur à y.

Type de descripteurs d’emplacement	Syntaxe	Description
Numéro de résidu unique	`x`	Désigne un résidu unique dans la séquence présentée.
Numéros de résidus délimitant un ensemble dans la séquence	`x..y`	Désigne une série continue de résidus délimitée par un résidu de début et un résidu de fin, ces deux résidus étant inclus dans la série.

Résidus situés avant le premier ou après le dernier numéro de résidu indiqué	`<x`
	`>x`
	`<x..y`
	`x..>y`
		Désigne une région qui comprend un résidu ou une série de résidus indiqués et qui s’étend au-delà d’un résidu indiqué. Les symboles “<” et “>” peuvent être employés à l’égard d’un résidu unique ou des numéros du résidu de début et de fin d’une série de résidus pour signaler qu’une caractéristique s’étend au-delà du numéro de résidu indiqué.

Site s’étendant entre deux numéros de résidus adjacents	`x^y`	Désigne un site entre deux résidus adjacents, par exemple le site d’un clivage endonucléolytique. Les numéros de position des résidus adjacents sont séparés par un caret (^). Les formats autorisés pour ce descripteur sont x^x+1 (par exemple 55^56), ou pour les nucléotides circulaires, x^1, où “x” est la longueur totale de la molécule, c’est-à-dire 1000^1 pour une molécule circulaire de longueur 1000.

68. 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.

a) Opérateur d'emplacement pour des nucléotides et des acides aminés :

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

b) Opérateur d'emplacement réservé aux nucléotides :

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

69. Les opérateurs d'emplacement assurant un rôle de jonction ou d'ordonnancement nécessitent au moins deux descripteurs d'emplacement séparés par des virgules. Les descripteurs d'emplacement concernant des sites situés entre deux résidus adjacents, c'est-à-dire x^y, ne peuvent être employés dans un emplacement de jonction ou d'ordonnancement. L'emploi de l'opérateur d'emplacement de jonction implique que les résidus désignés par les descripteurs d'emplacement sont physiquement mis en contact par des processus biologiques (par exemple, les exons qui contribuent à la caractéristique d'une région jouant un rôle de codage).

70. L'opérateur d'emplacement "complement" ne peut être employé que pour des nucléotides. Il peut être employé en combinaison soit avec "join" soit avec "order" dans le même emplacement. Les combinaisons de "join" et "order" ne sont pas autorisées dans un même emplacement.

71. Les exemples ci-après illustrent des emplacements de caractéristiques au sens des paragraphes 65 à 70 ci-dessus :

a) emplacements pour des nucléotides et des acides aminés :

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

b) emplacements réservés aux nucléotides :

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

72. 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

73. 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 86 et 87);

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

c) des séquences.

74. La section 6 de l'annexe I contient une liste complète des qualificateurs et la définition du format de leurs valeurs, le cas échéant, pour la clé de caractérisation de chaque nucléotide, et la section 8 contient la liste complète des qualificateurs pour la clé de caractérisation de chaque acide aminé.

75. Toute séquence prévue au paragraphe 6 qui est indiquée à titre de valeur d'un qualificateur doit être présentée de manière distincte dans le listage des séquences avec son propre numéro d'identification de séquence.

Qualificateurs de caractéristiques obligatoires

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

Éléments des qualificateurs

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

Élément	Description	Obligatoire/Facultatif
`INSDQualifier_name`	Nom du qualificateur (voir annexe I, sections 6 et 8)	Obligatoire
`INSDQualifier_value`	Valeur du qualificateur, le cas échéant, au format indiqué (voir annexe I, sections 6 et 8)	Obligatoire si indiqué (voir annexe I, sections 6 et 8)

78. Le qualificateur d'organisme, c'est-à-dire l'élément "organism" pour les séquences de nucléotides (voir annexe I, section 6) et "ORGANISM" pour les séquences d'acides aminés (voir annexe I, section 8) doit divulguer la source, c'est-à-dire l'organisme ou l'origine unique de la séquence qui est présentée. Les indications d'organisme doivent être choisies parmi les éléments d'une taxonomie.

79. Si la séquence existe à l'état naturel et qu'il existe une désignation de genre et d'espèce en latin pour l'organisme source, le qualificateur doit prendre cette désignation pour valeur. Il est possible d'indiquer le nom commun anglais le plus courant à l'aide du qualificateur "note" pour les séquences de nucléotides et "NOTE" pour les séquences d'acides aminés, mais ce nom ne doit pas être employé comme valeur du qualificateur d'organisme.

80. Les exemples suivants illustrent la source de séquences présentées conformément aux paragraphes 78 et 79 ci-dessus :

Exemple 1 : Source d'une séquence de nucléotides

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

Exemple 2 : Source d'une séquence de protéines

```
<INSDSeq_feature-table>
   <INSDFeature>
       <INSDFeature_key>SOURCE</INSDFeature_key>
       <INSDFeature_location>1..174</INSDFeature_location>
       <INSDFeature_quals>
           <INSDQualifier>
               <INSDQualifier_name>ORGANISM</INSDQualifier_name>
               <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
           </INSDQualifier>
           <INSDQualifier>
               <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
               <INSDQualifier_value>protein</INSDQualifier_value>
           </INSDQualifier>
       </INSDFeature_quals>
   </INSDFeature>
</INSDSeq_feature-table>
```

81. 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>
```

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

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

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

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

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

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

85. Le qualificateur "mol_type" pour les séquences de nucléotides (voir annexe I, section 6) et "MOL_TYPE" pour les séquences d'acides aminés (voir annexe I, section 8) doit divulguer le type de molécule représenté dans la séquence. Ces qualificateurs sont distincts de l'élément `INSDSeq_moltype` examiné au paragraphe 54 :

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

86. Le texte libre est un format de valeur autorisé pour certains qualificateurs (comme indiqué à l'annexe I). Il s'agit d'un texte descriptif qui se présente sous forme de segments de phrases, de préférence en anglais.

87. L'emploi du texte libre doit être limité à un petit nombre de termes brefs indispensables à la compréhension d'une caractéristique de la séquence. Pour chaque qualificateur, le texte libre ne peut compter plus de 1000 caractères.

Séquences de codage

88. La clé de caractérisation "CDS" peut servir à désigner des séquences de codage, c'est-à-dire des séquences de nucléotides correspondant à la séquence d'acides aminés dans une protéine et au codon d'arrêt. L'élément `INSDFeature_location` doit indiquer l'emplacement de la caractéristique "CDS", y compris le codon d'arrêt.

89. Les qualificateurs "transl_table" et "translation" peuvent être employés en association avec la clé de caractérisation "CDS" (voir annexe I). Si le qualificateur "transl_table" n'est pas employé, on présume que c'est le tableau de codes normalisés qui est appliqué (voir annexe I, section 9, tableau 5).

90. Toute séquence protéinique codée selon la séquence de codage et divulguée dans un qualificateur de type "translation" visé par le paragraphe 6 doit avoir son propre numéro d'identification de séquence et doit être présentée dans le listage des séquences. Le numéro d'identification de séquence attribué à la séquence protéinique doit figurer dans la valeur du qualificateur "protein_id" associé à la clé de caractérisation "CDS". Le qualificateur "ORGANISM" associé à la clé de caractérisation "SOURCE" de la séquence protéinique doit être identique à celui de sa séquence de codage, par exemple :

```
<INSDSeq_feature-table>
   <INSDFeature>
       <INSDFeature_key>CDS</INSDFeature_key>
       <INSDFeature_location>1..507</INSDFeature_location>
       <INSDFeature_quals>
           <INSDQualifier>
               <INSDQualifier_name>transl_table</INSDQualifier_name>
               <INSDQualifier_value>11</INSDQualifier_value>
           </INSDQualifier>
           <INSDQualifier>
               <INSDQualifier_name>translation</INSDQualifier_name>
               <INSDQualifier_value>
MLVHLERTTIMFDFSSLINLPLIWGLLIAIAVLLYILMDGFDLGIGILLPFAPSDKCRDHMISSIAPFWDGNETWLVLGGGGLFAA
FPLAYSILMPAFYIPIIIMLLGLIVRGVSFEFRFKAEGKYRRLWDYAFHFGSLGAAFCQGMILGAFIHGVEVNGRNFSGGQLM
               </INSDQualifier_value>
           </INSDQualifier>
           <INSDQualifier>
                  <INSDQualifier_name>protein_id</INSDQualifier_name>
               <INSDQualifier_value>89</INSDQualifier_value>
           </INSDQualifier>
       </INSDFeature_quals>
   </INSDFeature>
</INSDSeq_feature-table>
```

Variantes

91. Toute séquence primaire et toute variante de cette séquence, chacune d'elles étant divulguée par énumération de ses résidus et visée par le paragraphe 6, doit être présentée dans le listage des séquences avec son propre numéro d'identification de séquence.

92. Toute séquence de variante, divulguée uniquement par référence à un ou plusieurs suppressions, adjonctions ou remplacements effectués dans une séquence primaire figurant dans le listage des séquences, peut être présentée dans le listage des séquences. Si tel est le cas, cette séquence de variante :

a) peut être présentée par annotation de la séquence primaire, si elle comporte une ou plusieurs variations à un seul emplacement ou à plusieurs emplacements distincts et que les occurrences de ces variations sont indépendantes;

b) devrait être présentée en tant que séquence distincte avec son propre numéro d'identification de séquence, si elle comporte des variations à plusieurs emplacements distincts et que les occurrences de ces variations sont interdépendantes; et

c) doit être présentée en tant que séquence distincte avec son propre numéro d'identification de séquence, si elle comporte une séquence qui a été ajoutée ou remplacée et qui contient plus de 1000 résidus (voir le paragraphe 87).

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

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

94. L'annotation d'une séquence primaire effectuée pour une variante particulière doit comporter une clé de caractérisation et un qualificateur, conformément au tableau ci-dessus, et indiquer l'emplacement de la caractéristique. Toute suppression doit être représentée par une valeur de qualificateur vide. Tout résidu ajouté ou remplacé doit être indiqué dans le qualificateur "replace" ou "NOTE". La valeur du qualificateur "replace" ou "NOTE" est un texte libre qui ne doit pas dépasser 1000 caractères, conformément au paragraphe 87. Pour les séquences visées par le paragraphe 6 qui sont présentées à titre d'adjonction ou de remplacement de la valeur d'un qualificateur, se reporter au paragraphe 97. La valeur du qualificateur peut comporter un listage des résidus alternatifs pouvant être ajoutés ou remplacés.

95. Les symboles indiqués dans l'annexe I (voir respectivement les sections 1 à 4, tableaux 1 à 4) peuvent être employés pour représenter des résidus de variantes, selon les besoins. Si un résidu de variante est un résidu modifié qui ne figure pas dans les tableaux 2 ou 4 de l'annexe I, le nom complet non abrégé du résidu modifié doit être indiqué dans la valeur du qualificateur.

96. Les exemples ci-après illustrent la manière de présenter des variantes pour qu'elles soient conformes aux paragraphes 92 à 95 ci-dessus :

Exemple 1 : Clé de caractérisation "variation" pour un remplacement dans une séquence de nucléotides.
Une cytosine remplace le nucléotide défini à la position 413 de la séquence.

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

Exemple 2 : Clé de caractérisation "misc_difference" pour une suppression dans une séquence de nucléotides.
Le nucléotide à la position 413 de la séquence est supprimé.

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

Exemple 3 : Clé de caractérisation "misc_difference" pour une adjonction dans une séquence de nucléotides.
La séquence "atgccaaatat" est ajoutée entre les positions 100 et 101 de la séquence primaire.

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

Exemple 4 : Clé de caractérisation "VARIANT" pour un remplacement dans une séquence d'acides aminés.
L'acide aminé indiqué à la position 100 de la séquence peut être remplacé par I, A, F, Y, 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>
```

Exemple 5 : Clé de caractérisation “VARIANT” pour un remplacement dans une séquence d’acides aminés.
L’acide aminé indiqué à la position 100 de la séquence peut être remplacé par tout acide aminé sauf Lys, Arg ou His.

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

97. Toute séquence visée par le paragraphe 6 qui est présentée à titre d’adjonction ou de remplacement dans la valeur d’un qualificateur pour une annotation de séquence primaire doit aussi être présentée dans le listage des séquences avec son propre numéro d’identification de séquence.

[L’annexe I à la norme ST.26 suit]

ST.26 - ANNEX I

CONTROLLED VOCABULARY

Final Draft

Proposal presented by the SEQL Task Force for consideration and adoption at the CWS/4

TABLE OF CONTENTS

SECTION 1: LIST OF NUCLEOTIDES

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

Table 1: List of nucleotides

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

SECTION 2: LIST OF MODIFIED NUCLEOTIDES

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

Table 2: List of modified nucleotides

Abbreviation	Modified Nucleotide
ac4c	4-acetylcytidine
chm5u	5-(carboxyhydroxylmethyl)uridine
cm	2’-O-methylcytidine
cmnm5s2u	5-carboxymethylaminomethyl-2-thiouridine
cmnm5u	5-carboxymethylaminomethyluridine
d	dihydrouridine
fm	2’-O-methylpseudouridine
gal q	beta,D-galactosylqueosine
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
m5c	5-methylcytidine
m6a	N6-methyladenosine
m7g	7-methylguanosine
mam5u	5-methylaminomethyluridine
mam5s2u	5-methoxyaminomethyl-2-thiouridine

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

SECTION 3: LIST OF AMINO ACIDS

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

Table 3: List of amino acids

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

SECTION 4: LIST OF MODIFIED AND UNUSUAL AMINO ACIDS

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

Table 4: List of modified and unusual amino acids

Abbreviation	Modified or Unusual Amino acid
Aad	2-Aminoadipic acid
bAad	3-Aminoadipic acid
bAla	beta-Alanine, beta-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: FEATURES KEYS FOR NUCLEIC SEQUENCES

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

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

5.1.	Feature Key	attenuator
	Definition	1) region of DNA at which regulation of termination of transcription occurs, which controls the expression of some bacterial operons; 2) sequence segment located between the promoter and the first structural gene that causes partial termination of transcription
	Optional qualifiers	allele gene gene_synonym map note operon phenotype
	Organism scope	prokaryotes
	Molecule scope	DNA

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

5.3.	Feature Key	CAAT_signal
	Definition	CAAT box; part of a conserved sequence located about 75 bp up-stream of the start point of eukaryotic transcription units which may be involved in RNA polymerase binding; consensus=GG(C or T)CAATCT [1,2]
	Optional qualifiers	allele gene gene_synonym map note
	Organism scope	eukaryotes and eukaryotic viruses
	Molecule scope	DNA
	References	[1] Efstratiadis, A. et al. Cell 21, 653-668 (1980) [2] Nevins, J.R. "The pathway of eukaryotic mRNA formation" Ann Rev Biochem 52, 441-466 (1983)

5.4.	Feature Key	CDS
	Definition	coding sequence; sequence of nucleotides that corresponds with the sequence of amino acids in a protein (location includes stop codon); feature may include amino acid conceptual translation
	Optional qualifiers	allele artificial_location codon_start EC_number exception function gene gene_synonym map note number operon product protein_id pseudo pseudogene ribosomal_slippage standard_name translation transl_except transl_table trans_splicing
	Comment	codon_start qualifier has valid value of 1 or 2 or 3, indicating the offset at which the first complete codon of a coding feature can be found, relative to the first base of that feature; transl_table defines the genetic code table used if other than the Standard or universal genetic code table; genetic code exceptions outside the range of the specified tables are reported in transl_except qualifier; only one of the qualifiers translation and 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 amino acids

5.5.	Feature Key	centromere
	Definition	region of biological interest indentified 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

5.6.	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

5.7.	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
	Parent Key	CDS
	Organism scope	eukaryotes

5.8.	Feature Key	enhancer
	Definition	a cis-acting sequence that increases the utilization of (some) eukaryotic promoters, and can function in either orientation and in any location (upstream or downstream) relative to the promoter
	Optional qualifiers	allele bound_moiety gene gene_synonym map note standard_name
	Organism scope	eukaryotes and eukaryotic viruses

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

5.10.	Feature Key	GC_signal
	Definition	GC box; a conserved GC-rich region located upstream of the start point of eukaryotic transcription units which may occur in multiple copies or in either orientation; consensus=GGGCGG
	Optional qualifiers	allele gene gene_synonym map note
	Organism scope	eukaryotes and eukaryotic viruses

5.11.	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.

5.12.	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.

5.13.	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

5.14.	Feature Key	J_segment
	Definition	joining segment of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains
	Optional qualifiers	allele gene gene_synonym map note product pseudo pseudogene standard_name
	Parent Key	CDS
	Organism scope	eukaryotes

5.15.	Feature Key	LTR
	Definition	long terminal repeat, a sequence directly repeated at both ends of a defined sequence, of the sort typically found in retroviruses
	Optional qualifiers	allele function gene gene_synonym map note standard_name

5.16.	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

5.17.	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 feature key RBS is used for ribosome binding sites

5.18.	Feature Key	misc_difference
	Definition	featured sequence differs from the presented sequence at this location and cannot be described by any other Difference key (unsure, variation, or modified_base)
	Optional qualifiers	allele clone compare gene gene_synonym map note phenotype replace standard_name
	Comment	the misc_difference feature key should 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.

5.19.	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

5.20.	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 standard_name
	Molecule scope	DNA

5.21.	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

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

5.23.	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

5.24.	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

5.25.	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.

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

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

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

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

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

5.31.	Feature Key	polyA_signal
	Definition	recognition region necessary for endonuclease cleavage of an RNA transcript that is followed by polyadenylation; consensus=AATAAA [1]
	Optional qualifiers	allele gene gene_synonym map note
	Organism scope	eukaryotes and eukaryotic viruses
	References	[1] Proudfoot, N. and Brownlee, G.G. Nature 263, 211-214 (1976)

5.32.	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

5.33.	Feature Key	precursor_RNA
	Definition	any RNA species that is not yet the mature RNA product; may include 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

5.34.	Feature Key	prim_transcript
	Definition	primary (initial, unprocessed) transcript; includes 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

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

5.36.	Feature Key	promoter
	Definition	region on a DNA molecule involved in RNA polymerase binding to initiate transcription
	Optional qualifiers	allele bound_moiety function gene gene_synonym map note operon phenotype pseudo pseudogene standard_name
	Molecule scope	DNA

5.37.	Feature Key	protein_bind
	Definition	non-covalent protein binding site on nucleic acid
	Mandatory qualifiers	bound_moiety
	Optional qualifiers	allele function gene gene_synonym map note operon standard_name
	Comment	note that RBS is used for ribosome binding sites

5.38.	Feature Key	RBS
	Definition	ribosome binding site
	Optional qualifiers	allele gene gene_synonym map note standard_name
	References	[1] Shine, J. and Dalgarno, L. Proc Natl Acad Sci USA 71, 1342-1346 (1974) [2] Gold, L. et al. Ann Rev Microb 35, 365-403 (1981)
	Comment	in prokaryotes, known as the Shine-Dalgarno sequence: is located 5 to 9 bases upstream of the initiation codon; consensus GGAGGT [1,2]

5.39.	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

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

5.41.	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 standard_name
	Comment	rRNA sizes should be annotated with the product qualifier

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

5.43.	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

5.44.	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

5.45.	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

5.46.	Feature Key	STS
	Definition	sequence tagged site; short, single-copy DNA sequence that characterizes a mapping landmark on the genome and can be detected by PCR; a region of the genome can be mapped by determining the order of a series of STSs
	Optional qualifiers	allele gene gene_synonym map note standard_name
	Molecule scope	DNA
	Parent key	misc_binding
	Comment	STS location to include primer(s) in primer_bind key or primers

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

5.48.	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

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

5.50.	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

5.51.	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

5.52.	Feature Key	tRNA
	Definition	mature transfer RNA, a small RNA molecule (75-85 bases long) that mediates the translation of a nucleic acid sequence into an amino acid sequence
	Optional qualifiers	allele anticodon function gene gene_synonym map note product pseudo pseudogene standard_name trans_splicing

5.53.	Feature Key	unsure
	Definition	author is unsure of exact sequence in this region
	Optional qualifiers	allele compare gene gene_synonym map note replace
	Comment	use the replace qualifier to annotate a deletion, insertion, or substitution.

5.54.	Feature Key	V_region
	Definition	variable region of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; codes for the variable amino terminal portion; can be composed of V_segments, D_segments, N_regions, and J_segments
	Optional qualifiers	allele gene gene_synonym map note product pseudo pseudogene standard_name
	Parent Key	CDS
	Organism scope	eukaryotes

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

5.56.	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; variability arising as a result of genetic manipulation (e.g. site directed mutagenesis) should be described with the misc_difference feature; use the replace qualifier to annotate a deletion, insertion, or substitution

5.57.	Feature Key	3'UTR
	Definition	region at the 3' end of a mature transcript (following the 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>.

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

5.59.	Feature Key	-10_signal
	Definition	Pribnow box; a conserved region about 10 bp upstream of the start-point of bacterial transcription units which may be involved in binding RNA polymerase; consensus=TAtAaT [1,2,3,4]
	Optional qualifiers	allele gene gene_synonym map note operon standard_name
	Organism scope	prokaryotes
	Molecule scope	DNA
	References	[1] Schaller, H., Gray, C., and Hermann, K. Proc Natl Acad Sci USA 72, 737-741 (1974) [2] Pribnow, D. Proc Natl Acad Sci USA 72, 784-788 (1974) [3] Hawley, D.K. and McClure, W.R. "Compilation and analysis of Escherichia coli promoter DNA sequences" Nucl Acid Res 11, 2237-2255 (1983) [4] Rosenberg, M. and Court, D. "Regulatory sequences involved in the promotion and termination of RNA transcription" Ann Rev Genet 13, 319-353 (1979)

5.60.	Feature Key	-35_signal
	Definition	a conserved hexamer about 35 bp upstream of the start.point of bacterial transcription units; consensus=TTGACa or TGTTGACA
	Optional qualifiers	allele gene gene_synonym map note operon standard_name
	Organism scope	prokaryotes
	Molecule scope	DNA
	References	[1] Takanami, M., et al. Nature 260, 297-302 (1976) [2] Moran, C.P., Jr., et al. Molec Gen Genet 186, 339-346 (1982) [3] Maniatis, T., et al. Cell 5, 109-113 (1975)

SECTION 6: DESCRIPTION OF QUALIFIERS FOR NUCLEIC SEQUENCES

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

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

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

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

6.3.	Qualifier	bound_moiety
	Definition	name of the molecule/complex that may bind to the given feature
	Value format	free text
	Example	<INSDQualifier_value>GAL4</INSDQualifier_value>
	Comment	Multiple bound_moiety qualifiers are legal on "promoter" and "enhancer" features. A single bound_moiety qualifier is legal on the "misc_binding", "oriT" and "protein_bind" features.

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

6.5.	Qualifier	cell_type
	Definition	cell type from which the sequence was obtained
	Value format	free text
	Example	<INSDQualifier_value>leukocyte</INSDQualifier_value>

6.6.	Qualifier	chromosome
	Definition	chromosome (e.g. Chromosome number) from which the sequence was obtained
	Value format	free text
	Example	<INSDQualifier_value>1</INSDQualifier_value> <INSDQualifier_value>X</INSDQualifier_value>

6.7.	Qualifier	clone
	Definition	clone from which the sequence was obtained
	Value format	free text
	Example	<INSDQualifier_value>lambda-hIL7.3</INSDQualifier_value>
	Comment	not more than one clone should be specified for a given source feature; where the sequence was obtained from multiple clones it may be further described in the feature table using the feature key misc_feature and a note qualifier to specify the multiple clones.

6.8.	Qualifier	clone_lib
	Definition	clone library from which the sequence was obtained
	Value format	free text
	Example	<INSDQualifier_value>lambda-hIL7</INSDQualifier_value>

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

6.10.	Qualifier	collected_by
	Definition	name of persons or institute who collected the specimen
	Value format	free text
	Example	<INSDQualifier_value>Dan Janzen</INSDQualifier_value>

6.11.	Qualifier	collection_date
	Definition	date that the specimen was collected
	Value format	DD-Mmm-YYYY, Mmm-YYYY or YYYY
	Example	<INSDQualifier_value>21-Oct-1952</INSDQualifier_value> <INSDQualifier_value>Oct-1952</INSDQualifier_value> <INSDQualifier_value>1952</INSDQualifier_value>
	Comment	full date format DD-Mmm-YYYY is preferred; where day and/or month of collection is not known either "Mmm-YYYY" or "YYYY" can be used; three-letter month abbreviation can be one of the following: Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec.

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

6.13.	Qualifier	cultivar
	Definition	cultivar (cultivated variety) of plant from which sequence was obtained
	Value format	free text
	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.

6.14.	Qualifier	dev_stage
	Definition	if the sequence was obtained from an organism in a specific developmental stage, it is specified with this qualifier
	Value format	free text
	Example	<INSDQualifier_value>fourth instar larva</INSDQualifier_value>

6.15.	Qualifier	direction
	Definition	direction of DNA replication
	Value format	left, right, or both where left indicates toward the 5’ end of the sequence (as presented) and right indicates toward the 3’ end
	Example	<INSDQualifier_value>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. The values are case-insensitive, i.e. both "RIGHT" and "right" are valid.

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

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

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

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

6.20.	Qualifier	frequency
	Definition	frequency of the occurrence of a feature
	Value format	free text representing the proportion of a population carrying the feature expressed as a fraction
	Example	<INSDQualifier_value>23/108</INSDQualifier_value> <INSDQualifier_value>1 in 12</INSDQualifier_value> <INSDQualifier_value>0.85</INSDQualifier_value>

6.21.	Qualifier	function
	Definition	function attributed to a sequence
	Value format	free text
	Example	<INSDQualifier_value>essential for recognition of cofactor </INSDQualifier_value>
	Comment	The function qualifier is used when the gene name and/or product name do not convey the function attributable to a sequence.

6.22.	Qualifier	gene
	Definition	symbol of the gene corresponding to a sequence region
	Value format	free text
	Example	<INSDQualifier_value>ilvE</INSDQualifier_value>
	Comment	Use gene qualifier to provide the gene symbol; use standard_name qualifier to provide the full gene name.

6.23.	Qualifier	gene_synonym
	Definition	synonymous, replaced, obsolete or former gene symbol
	Value format	free text
	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 used, a primary gene symbol must always be indicated in a gene qualifier

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

6.25.	Qualifier	haplogroup
	Definition	name for a group of similar haplotypes that share some sequence variation.

		Haplogroups are often used to track migration of population groups.
	Value format	free text
	Example	<INSDQualifier_value>H*</INSDQualifier_value>

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

6.27.	Qualifier	host
	Definition	natural (as opposed to laboratory) host to the organism from which sequenced molecule was obtained
	Value format	free text
	Example	<INSDQualifier_value>Homo sapiens</INSDQualifier_value> <INSDQualifier_value>Homo sapiens 12 year old girl</INSDQualifier_value> <INSDQualifier_value>Rhizobium NGR234</INSDQualifier_value>

6.28.	Qualifier	identified_by
	Definition	name of the expert who identified the specimen taxonomically
	Value format	free text
	Example	<INSDQualifier_value>John Burns</INSDQualifier_value>

6.29.	Qualifier	isolate
	Definition	individual isolate from which the sequence was obtained
	Value format	free text
	Example	<INSDQualifier_value>Patient #152</INSDQualifier_value> <INSDQualifier_value>DGGE band PSBAC-13</INSDQualifier_value>

6.30.	Qualifier	isolation_source
	Definition	describes the physical, environmental and/or local geographical source of the biological sample from which the sequence was derived
	Value format	free text
	Examples	<INSDQualifier_value>rumen isolates from standard Pelleted ration-fed steer #67</INSDQualifier_value> <INSDQualifier_value>permanent Antarctic sea ice</INSDQualifier_value> <INSDQualifier_value>denitrifying activated sludge from carbon_limited continuous reactor</INSDQualifier_value>
	Comment	used only with the source feature key; source feature keys containing an environmental_sample qualifier should also contain an isolation_source qualifier

6.31.	Qualifier	lab_host
	Definition	scientific name of the laboratory host used to propagate the source organism from

		which the sequenced molecule was obtained
	Value format	free text
	Example	<INSDQualifier_value>Gallus gallus</INSDQualifier_value> <INSDQualifier_value>Gallus gallus embryo</INSDQualifier_value> <INSDQualifier_value>Escherichia coli strain DH5 alpha</INSDQualifier_value> <INSDQualifier_value>Homo sapiens HeLa cells</INSDQualifier_value>
	Comment	the full binomial scientific name of the host organism should be used when known; extra conditional information relating to the host may also be included
6.32.	Qualifier	lat_lon
	Definition	geographical coordinates of the location where the specimen was collected
	Value format	free text - degrees latitude and longitude in format "d[d.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>
6.33.	Qualifier	macronuclear
	Definition	if the sequence shown is DNA and from an organism which undergoes chromosomal differentiation between macronuclear and micronuclear stages, this qualifier is used to denote that the sequence is from macronuclear DNA
	Value format	none
6.34.	Qualifier	map
	Definition	genomic map position of feature
	Value format	free text
	Example	<INSDQualifier_value>8q12-13</INSDQualifier_value>
6.35.	Qualifier	mating_type
	Definition	mating type of the organism from which the sequence was obtained; mating type is used for prokaryotes, and for eukaryotes that undergo meiosis without sexually dimorphic gametes
	Value format	free text
	Examples	<INSDQualifier_value>MAT-1</INSDQualifier_value> <INSDQualifier_value>plus</INSDQualifier_value> <INSDQualifier_value>-</INSDQualifier_value> <INSDQualifier_value>odd</INSDQualifier_value> <INSDQualifier_value>even</INSDQualifier_value>"
	Comment	mating_type qualifier values male and female are valid in the prokaryotes, but not in the eukaryotes; for more information, see the entry for the sex qualifier.

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

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

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

6.39.	Qualifier	ncRNA_class
	Definition	a structured description of the classification of the non-coding RNA described by the ncRNA parent key
	Value format	TYPE where TYPE is one of the following controlled vocabulary terms or phrases: antisense_RNA autocatalytically_spliced_intron ribozyme hammerhead_ribozyme lncRNA RNase_P_RNA RNase_MRP_RNA telomerase_RNA guide_RNA rasiRNA scRNA siRNA miRNA piRNA snoRNA snRNA SRP_RNA" vault_RNA Y_RNA other
	Example	<INSDQualifier_value>autocatalytically_spliced_intron </INSDQualifier_value> <INSDQualifier_value>siRNA</INSDQualifier_value> <INSDQualifier_value>scRNA</INSDQualifier_value> <INSDQualifier_value>other</INSDQualifier_value>
	Comment	specific ncRNA types not yet in the ncRNA_class controlled vocabulary can be annotated by entering "other" as the ncRNA_class qualifier value, and providing a brief explanation of novel ncRNA_class in a note qualifier

6.40.	Qualifier	note
	Definition	any comment or additional information
	Value format	free text
	Example	<INSDQualifier_value>A comment about the feature</INSDQualifier_value>

6.41.	Qualifier	number
	Definition	a number to indicate the order of genetic elements (e.g. exons or introns) in the 5’ to 3’ direction
	Value format	free text (with no whitespace characters)
	Example	<INSDQualifier_value>4</INSDQualifier_value> <INSDQualifier_value>6B</INSDQualifier_value>
	Comment	text limited to integers, letters or combination of integers and/or letters represented as a data value that contains no whitespace characters; any additional terms should be included in a standard_name qualifier. Example: a number qualifier with a value of 2A and a standard_name qualifier with a value of long

6.42.	Qualifier	operon
	Definition	name of the group of contiguous genes transcribed into a single transcript to which that feature belongs
	Value format	free text
	Example	<INSDQualifier_value>lac</INSDQualifier_value>
	Comment	valid only on Prokaryota-specific features

6.43.	Qualifier	organelle
	Definition	type of membrane-bound intracellular structure from which the sequence was obtained
	Value format	One of the following controlled vocabulary terms and phrases: chromatophore hyrogenosome mitochondrion nucleomorph plastid mitochondrion:kinetoplast plastid:chloroplast plastid:apicoplast plastid:chromoplast plastid:cyanelle plastid:leucoplast plastid:proplastid,
	Examples	<INSDQualifier_value>chromatophore</INSDQualifier_value> <INSDQualifier_value>hydrogenosome</INSDQualifier_value> <INSDQualifier_value>mitochondrion</INSDQualifier_value> <INSDQualifier_value>nucleomorph</INSDQualifier_value> <INSDQualifier_value>plastid</INSDQualifier_value> <INSDQualifier_value>mitochondrion:kinetoplast</INSDQualifier_value> <INSDQualifier_value>plastid:chloroplast</INSDQualifier_value> <INSDQualifier_value>plastid:apicoplast</INSDQualifier_value> <INSDQualifier_value>plastid:chromoplast</INSDQualifier_value> <INSDQualifier_value>plastid:cyanelle</INSDQualifier_value> <INSDQualifier_value>plastid:leucoplast</INSDQualifier_value> <INSDQualifier_value>plastid:proplastid</INSDQualifier_value>

6.44.	Qualifier	organism
	Definition	scientific name of the organism that provided the sequenced genetic material, if known, or the available taxonomic information if the organism is unclassified; or an indication that the sequence is a synthetic construct
	Value format	free text
	Example	<INSDQualifier_value>Homo sapiens</INSDQualifier_value>

6.45.	Qualifier	PCR_primers
	Definition	PCR primers that were used to amplify the sequence. A single /PCR_primers qualifier should contain all the primers used for a single PCR reaction. If multiple forward or reverse primers are present in a single PCR reaction, multiple sets of fwd_name/fwd_seq or rev_name/rev_seq values will be present
	Value format	[fwd_name: XXX1,]fwd_seq: xxxxx1,[fwd_name: XXX2,]fwd_seq: xxxxx2, [rev_name: YYY1,]rev_seq: yyyyy1,[rev_name: YYY2,]rev_seq: yyyyy2</INSDQualifier_value>
	Example	<INSDQualifier_value>fwd_name: 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 should be presented in 5’>3’ order. The sequences should be given in the symbols from Section 1 of this Annex, except for the modified bases; those must be enclosed within angle brackets < >. In XML, the angle brackets < and > must be substituted with < and > since they are reserved characters in XML.

6.46.	Qualifier	phenotype
	Definition	phenotype conferred by the feature, where phenotype is defined as a physical, biochemical or behavioural characteristic or set of characteristics
	Value format	free text
	Example	<INSDQualifier_value>erythromycin resistance</INSDQualifier_value>

6.47.	Qualifier	plasmid
	Definition	name of naturally occurring plasmid from which the sequence was obtained, where plasmid is defined as an independently replicating genetic unit that cannot be described by chromosome or segment qualifiers
	Value format	free text
	Example	<INSDQualifier_value>pC589</INSDQualifier_value>

6.48.	Qualifier	pop_variant
	Definition	name of subpopulation or phenotype of the sample from which the sequence was derived
	Value format	free text
	Example	<INSDQualifier_value>pop1</INSDQualifier_value> <INSDQualifier_value>Bear Paw</INSDQualifier_value>

6.49.	Qualifier	product
	Definition	name of the product associated with the feature, e.g. the mRNA of an mRNA feature, the polypeptide of a CDS, the mature peptide of a mat_peptide, etc.
	Value format	free text
	Example	<INSDQualifier_value>trypsinogen</INSDQualifier_value> (when qualifier appears in CDS feature) <INSDQualifier_value>trypsin</INSDQualifier_value> (when qualifier appears in mat_peptide feature) <INSDQualifier_value>XYZ neural-specific transcript</INSDQualifier_value> (when qualifier appears in mRNA feature)

6.50.	Qualifier	protein_id
	Definition	protein sequence identification number, an integer used in a sequence listing to designate the protein sequence encoded by the coding sequence identified in the corresponding CDS feature key
	Value format	an integer greater than zero
	Example	<INSDQualifier_value>89</INSDQualifier_value>

6.51.	Qualifier	proviral
	Definition	this qualifier is used to flag sequence obtained from a virus or phage that is integrated into the genome of another organism
	Value format	none

6.52.	Qualifier	pseudo
	Definition	indicates that this feature is a non-functional version of the element named by the feature key
	Value format	none
	Comment	The qualifier pseudo should be used to describe non-functional genes that are not formally described as pseudogenes, e.g. CDS has no translation due to other reasons than pseudogenisation events. Other reasons may include sequencing or assembly errors. In order to annotate pseudogenes the qualifier pseudogene must be used, indicating the TYPE of pseudogene.

6.53.	Qualifier	pseudogene
	Definition	indicates that this feature is a pseudogene of the element named by the feature key
	Value format	TYPE where TYPE is one of the following controlled vocabulary terms or phrases: processed unprocessed unitary allelic unknown
	Example	<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 mutation. 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 pseudogenisation.

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

6.55.	Qualifier	replace
	Definition	indicates that the sequence identified in a feature’s location is replaced by the sequence shown in the qualifier’s value; if no sequence (i.e., no value) is contained within the qualifier, this indicates a deletion
	Value format	free text
	Example	<INSDQualifier_value>a</INSDQualifier_value> <INSDQualifier_value></INSDQualifier_value> - for a deletion

6.56.	Qualifier	ribosomal_slippage
	Definition	during protein translation, certain sequences can program ribosomes to change to an alternative reading frame by a mechanism known as ribosomal slippage
	Value format	none
	Comment	a join operator, e.g.: [join(486..1784,1787..4810)] should be used in the CDS spans to indicate the location of ribosomal_slippage

6.57.	Qualifier	rpt_family
	Definition	type of repeated sequence; "Alu" or "Kpn", for example
	Value format	free text
	Example	<INSDQualifier_value>Alu</INSDQualifier_value>

6.58.	Qualifier	rpt_type
	Definition	organization of repeated sequence
	Value format	One of the following controlled vocabulary terms: tandem inverted flanking terminal direct dispersed other
	Example	<INSDQualifier_value>INVERTED</INSDQualifier_value>
	Comment	the values are case-insensitive, i.e. both "INVERTED" and "inverted" are valid; Definitions of the values: tandem - a repeat that exists adjacent to another in the same orientation; inverted - a repeat which occurs as part of as set (normally a part) organized in the reverse orientation; flanking - a repeat lying outside the sequence for which it has functional significance (eg. transposon insertion target sites); terminal - a repeat at the ends of and within the sequence for which it has functional significance (eg. transposon LTRs); direct - a repeat that exists not always adjacent but is in the same orientation; dispersed - a repeat that is found dispersed throughout the genome; other - a repeat exhibiting important attributes that cannot be described by other values.

6.59.	Qualifier	rpt_unit_range
	Definition	location (range) of a repeating unit
	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.

6.60.	Qualifier	rpt_unit_seq
	Definition	identity of a repeat sequence
	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

6.61.	Qualifier	satellite
	Definition	identifier for a satellite DNA marker, compose of many tandem repeats (identical or related) of a short basic repeated unit
	Value format	<satellite_type>[:<class>][<identifier>] - where <satellite_type> is one of the following: satellite; microsatellite; minisatellite
	Example	<INSDQualifier_value>satellite: S1a</INSDQualifier_value> <INSDQualifier_value>satellite: alpha</INSDQualifier_value> <INSDQualifier_value>satellite: gamma III</INSDQualifier_value> <INSDQualifier_value>microsatellite: DC130</INSDQualifier_value>
	Comment	many satellites have base composition or other properties that differ from those of the rest of the genome that allows them to be identified.

6.62.	Qualifier	segment
	Definition	name of viral or phage segment sequenced
	Value format	free text
	Example	<INSDQualifier_value>6</INSDQualifier_value>

6.63.	Qualifier	serotype
	Definition	serological variety of a species characterized by its antigenic properties
	Value format	free text
	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".

6.64.	Qualifier	serovar
	Definition	serological variety of a species (usually a prokaryote) characterized by its antigenic properties
	Value format	free text
	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".

6.65.	Qualifier	sex
	Definition	sex of the organism from which the sequence was obtained; sex is used for eukaryotic organisms that undergo meiosis and have sexually dimorphic gametes
	Value format	free text
	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

6.66.	Qualifier	standard_name
	Definition	accepted standard name for this feature
	Value format	free text
	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).

6.67.	Qualifier	strain
	Definition	strain from which sequence was obtained
	Value format	free text
	Example	<INSDQualifier_value>BALB/c</INSDQualifier_value>
	Comment	entries including strain qualifier must not include the environmental_sample qualifier

6.68.	Qualifier	sub_clone
	Definition	sub-clone from which sequence was obtained
	Value format	free text
	Example	<INSDQualifier_value>lambda-hIL7.20g</INSDQualifier_value>
	Comment	not more than one sub_clone should be specified for a given source feature; to indicate that the sequence was obtained from multiple sub_clones, multiple source features should be given

6.69.	Qualifier	sub_species
	Definition	name of sub-species of organism from which sequence was obtained
	Value format	free text
	Example	<INSDQualifier_value>lactis</INSDQualifier_value>

6.70.	Qualifier	sub_strain
	Definition	name or identifier of a genetically or otherwise modified strain from which sequence was obtained, derived from a parental strain (which should be annotated in the strain qualifier). sub_strain from which sequence was obtained
	Value format	free text
	Example	<INSDQualifier_value>abis</INSDQualifier_value>
	Comment	If the parental strain is not given, this 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

6.71.	Qualifier	tag_peptide
	Definition	base location encoding the polypeptide for proteolysis tag of tmRNA and its termination codon
	Value format	<base_range> - where <base_range> provides the first and last base (separated by two dots) of the location for the proteolysis tag
	Example	<INSDQualifier_value>90..122</INSDQualifier_value>
	Comment	it is recommended that the amino acid sequence corresponding to the tag_peptide be annotated by describing a 5’ partial CDS feature; e.g. CDS with a location of <90..122

6.72.	Qualifier	tissue_lib
	Definition	tissue library from which sequence was obtained
	Value format	free text
	Example	<INSDQualifier_value>tissue library 772</INSDQualifier_value>

6.73.	Qualifier	tissue_type
	Definition	tissue type from which the sequence was obtained
	Value format	free text
	Example	<INSDQualifier_value>liver</INSDQualifier_value>

6.74.	Qualifier	transl_except
	Definition	translational exception: single codon the translation of which does not conform to genetic code defined by organism or transl_table.
	Value format	(pos:location,aa:<amino_acid>) where <amino_acid> is the amino acid coded by the codon at the base_range position
	Example	<INSDQualifier_value>(pos:213..215,aa:Trp) </INSDQualifier_value> <INSDQualifier_value>(pos:462..464,aa:OTHER) </INSDQualifier_value> <INSDQualifier_value>(pos:1017,aa:TERM) </INSDQualifier_value> <INSDQualifier_value>(pos:2000..2001,aa:TERM) </INSDQualifier_value> <INSDQualifier_value>(pos:X22222:15..17,aa:Ala) </INSDQualifier_value>
	Comment	if the amino acid is not one of the specific amino acids listed in Section 3 of this Annex, use OTHER as <amino_acid> and provide the name of the unusual amino acid in a note qualifier; for modified amino-acid selenocysteine use three letter code 'Sec' (one letter code 'U' 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'.

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

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

6.77.	Qualifier	translation
	Definition	one-letter abbreviated amino acid sequence derived from either the standard (or universal) genetic code or the table as specified in a transl_table qualifier and as determined by an exception in the transl_except qualifier
	Value format	contiguous string of one-letter amino acid abbreviations from Section 3 of this Annex, "X" is to be used for AA exceptions.
	Example	<INSDQualifier_value>MASTFPPWYRGCASTPSLKGLIMCTW</INSDQualifier_value>
	Comment	to be used with CDS feature only; must be accompanied by protein_id qualifier when the translation product contains four or more 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.

6.78.	Qualifier	variety
	Definition	variety (= varietas, a formal Linnaean rank) of organism from which sequence was derived.
	Value format	free text
	Example	<INSDQualifier_value>insularis</INSDQualifier_value>
	Comment	use the cultivar qualifier for cultivated plant varieties, i.e., products of artificial selection; varieties other than plant and fungal variatas should be annotated via a note qualifier, e.g. with the value <INSDQualifier_value>breed:Cukorova</INSDQualifier_value>

SECTION 7: FEATURE KEYS FOR AMINO ACID SEQUENCES

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

7.1.	Feature Key	ACT_SITE
	Definition	Amino acid(s) involved in the activity of an enzyme
	Optional qualifiers	NOTE
	Comment	Each amino acid resdidue of the active site should be annotated separately with the ACT_SITE feature key. The corresponding amino acid residue number should 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. 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. The type of linkage (C-, N- or O-linked) to the protein is indicated in the “NOTE” qualifier. Examples of values used in the “NOTE” qualifier: O-linked (GlcNAc); C-linked (Man); N-linked (GlcNAc...); and O-linked (Glc...).

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

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

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

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

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

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

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

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

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

7.17.	Feature Key	METAL
	Definition	Binding site for a metal ion.
	Mandatory qualifiers	NOTE
	Comment	The NOTE qualifier indicates the nature of the metal. Examples of values for the “NOTE” qualifier: “Iron; catalytic” 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 Table 4 of this Annex, the abbreviation may be used in place of the the full name. Examples of values for the “NOTE” qualifier: “N-acetylalanine”; “3-Hyp”; and “MeLys” or “N-6-methyllysine"

7.19.	Feature Key	MOTIF
	Definition	Short (up to 20 amino acids) sequence motif of biological interest
	Optional qualifiers	NOTE

7.20.	Feature Key	MUTAGEN
	Definition	Site which has been experimentally altered by mutagenesis
	Optional qualifiers	NOTE

7.21.	Feature Key	NON_STD
	Definition	Non-standard amino acid
	Optional qualifiers	NOTE
	Comment	This key describes the occurrence of non-standard amino acids selenocysteine (U) and pyrrolysine (O) in the amino acid sequence.

7.22.	Feature Key	NON_TER
	Definition	The residue at an extremity of the sequence is not the terminal residue
	Optional qualifiers	NOTE
	Comment	If applied to position 1, this means that the first position is not the N-terminus of the complete molecule. If applied to the last position, it means that this position is not the C-terminus of the complete molecule.

7.23.	Feature Key	NP_BIND
	Definition	Extent of a nucleotide phosphate-binding region
	Mandatory qualifiers	NOTE
	Comment	The nature of the nucleotide phosphate is indicated in the NOTE qualifier. Examples of values for the “NOTE” qualifier: “ATP” and “FAD”.

7.24.	Feature Key	PEPTIDE
	Definition	Extent of a released active peptide
	Optional qualifiers	NOTE

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

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

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

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

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

7.30.	Feature Key	SOURCE
	Definition	Identifies the source of the sequence; this key is mandatory; every sequence will have a single SOURCE feature spanning the entire sequence
	Mandatory qualifiers	MOL_TYPE ORGANISM
	Optional qualifiers	NOTE

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

7.32.	Feature Key	TOPO_DOM
	Definition	Topological domain
	Optional qualifiers	NOTE

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

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

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

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

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

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

7.39.	Feature Key	ZN_FING
	Definition	Extent of a zinc finger region
	Mandatory qualifiers	NOTE
	Comment	The type of zinc finger is indicated in the NOTE qualifier. For example: “GATA-type” and “NR C4-type”

SECTION 8: QUALIFIERS FOR AMINO ACID SEQUENCES

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

8.1.	Qualifier	MOL_TYPE
	Definition	In vivo molecule type of sequence
	Value format	protein
	Example	<INSDQualifier_value>protein</INSDQualifier_value>
	Comment	The "MOL_TYPE" qualifier is mandatory on the SOURCE feature key.

8.2.	Qualifier	NOTE
	Definition	Any comment or additional information
	Value format	free text
	Example	<INSDQualifier_value>Heme (covalent)</INSDQualifier_value>
	Comment	The “NOTE” qualifier is mandatory for the feature keys: BINDING; CARBOHYD; CROSSLNK; DISULFID; DNA_BIND; DOMAIN; LIPID; METAL; MOD_RES; NP_BIND and ZN_FING

8.3.	Qualifier	ORGANISM
	Definition	Scientific name of the organism that provided the peptide
	Value format	free text
	Example	<INSDQualifier_value>Homo sapiens</INSDQualifier_value>
	Comment	The “ORGANISM” qualifier is mandatory for the SOURCE feature key.

SECTION 9: GENETIC CODES TABLES

Table 5 reproduces Genetic Code Tables to be used for translating coding sequences. The value for the trans_table qualifier is the number assigned to the corresponding genetic code table. Where a CDS feature is described with a translation qualifier but not a transl_table qualifier, the 1 - Standard Code is used by default for translation. (Note: Genetic code tables 7, 8, and 17 to 20 do not exist, therefore these numbers do not appear in Table 5.)

Table 5: Genetic Code Tables

1 - Standard Code
AAs = FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG Starts = ---M---------------M---------------M---------------------------- Base1 = tttttttttttttttcccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggg Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggg Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
2 - Vertebrate Mitochondrial Code
AAs = FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSS**VVVVAAAADDEEGGGG Starts = --------------------------------MMMM---------------M------------ Base1 = tttttttttttttttcccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggg Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggg Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
3 - Yeast Mitochondrial Code
AAs = FFLLSSSSYY**CCWWTTTTPPPPHHQQRRRRIIMMTTTTNNKKSSRRVVVVAAAADDEEGGGG Starts = ----------------------------------MM---------------------------- Base1 = tttttttttttttttcccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggg Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggg Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
4 - Mold, Protozoan, Coelenterate Mitochondrial Code & Mycoplasma/Spiroplasma Code
AAs = FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG Starts = --MM---------------M------------MMMM---------------M------------ Base1 = tttttttttttttttcccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggg Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggg Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
5 - Invertebrate Mitochondrial Code
AAs = FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSSSSVVVVAAAADDEEGGGG Starts = ---M------------------------------MMMM---------------M------------ Base1 = tttttttttttttttcccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggg Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggg Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
6 - Ciliate, Dasycladacean and Hexamita Nuclear Code
AAs = FFLLSSSSYYQQCC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG Starts = ----------------------------------M---------------------------- Base1 = tttttttttttttttcccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggg Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggg Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
9 - Echinoderm and Flatworm Mitochondrial Code
AAs = FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNNKSSSSVVVVAAAADDEEGGGG Starts = ----------------------------------M---------------M------------ Base1 = tttttttttttttttcccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggg Base2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggg Base3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag

10 - Euplotid Nuclear Code

```
  AAs  =  FFLLSSSSYY**CCCWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts =  ----------------------------------M----------------------------
Base1  =  tttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggg
Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

11 - Bacterial and Plant Plastid Code

```
  AAs  =  FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts =  ---M---------------M------------MMMM---------------M------------
Base1  =  tttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggg
Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

12 - Alternative Yeast Nuclear Code

```
  AAs  =  FFLLSSSSYY**CC*WLLLSPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts =  -------------------M---------------M----------------------------
Base1  =  tttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggg
Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

13- Ascidian Mitochondrial Code

```
  AAs  =  FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSSGGVVVVAAAADDEEGGGG
Starts =  ---M------------------------------MM---------------M------------
Base1  =  tttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggg
Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

14 - Alternative Flatworm Mitochondrial Code

```
  AAs  =  FFLLSSSSYYY*CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNNKSSSSVVVVAAAADDEEGGGG
Starts =  ----------------------------------M----------------------------
Base1  =  tttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggg
Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

15 - Blepharisma Nuclear Code

```
  AAs  =  FFLLSSSSYY*QCC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts =  ----------------------------------M----------------------------
Base1  =  tttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggg
Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

16 - Chlorophycean Mitochondrial Code

```
  AAs  =  FFLLSSSSYY*LCC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts =  ----------------------------------M----------------------------
Base1  =  tttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggg
Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

21 - Trematode Mitochondrial Code

```
  AAs  =  FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNNKSSSSVVVVAAAADDEEGGGG
Starts =  ----------------------------------M---------------M------------
Base1  =  tttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggg
Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

22 - Scenedesmus obliquus Mitochondrial Code

```
  AAs  =  FFLLSS*SYY*LCC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts =  ----------------------------------M----------------------------
Base1  =  ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggg
Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

23 - Thraustochytrium Mitochondrial Code

```
  AAs  =  FF*LSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts =  --------------------------------M--M---------------M------------
Base1  =  ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggg
Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

24 - Pterobranchia Mitochondrial Code

```
  AAs  =  FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSSKVVVVAAAADDEEGGGG
Starts =  ---M--------------M--------------M--------------M------------
Base1  =  ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggg
Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

25 - Candidate Division SR1 and Gracilibacteria Code

```
  AAs  = FFLLSSSSYY**CCGWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = ---M---------------M---------------M----------------------------
Base1  = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggg
Base2  = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggg
Base3  = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

[Annex II to ST.26 follows]

ST.26 - ANNEX II

DOCUMENT TYPE DEFINITION FOR SEQUENCE LISTING (DTD)

Final Draft

Proposal presented by the SEQL Task Force for consideration and adoption at the CWS/4

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

This entity may be identified by the PUBLIC identifier:
********************************************************************************
PUBLIC "-//WIPO//DTD SEQUENCE LISTING 1.0//EN" "ST26SequenceListing_V1_0.dtd"
********************************************************************************

************

* PUBLIC DTD URL

* http://www.wipo.int/standards/DTD/ST26SequenceListing_V1_0.dtd
********************************************************************************

Recommended Standard for the presentation of nucleotide and amino acid sequence listings
using XML (eXtensible Markup Language)

********************************************************************************
* CONTACTS
********************************************************************************

xml.standards@wipo.int


Date draft created: 2014-03-11

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

********************************************************************************
* REVISION HISTORY
********************************************************************************
2014-03-11

Final draft for adoption.
********************************************************************************

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

<!--The elements ApplicantName and InventorName are optional in this DTD to facilitate
the conversion between various encoding schemes-->
<!ATTLIST ST26SequenceListing
               dtdVersion  CDATA    #REQUIRED
               fileName  CDATA    #IMPLIED
               softwareName  CDATA    #IMPLIED
               softwareVersion  CDATA    #IMPLIED
               productionDate  CDATA    #IMPLIED  >

<!--ApplicantFileReference
```

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

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

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

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

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

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

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

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

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

<!--SequenceData
Data for individual Sequence.
```

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

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

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

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

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

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

The INSD part is subset of the INSD DTD v1.4 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.
-->
<!ELEMENT INSDSeq  (INSDSeq_length,INSDSeq_moltype,INSDSeq_division,
               INSDSeq_other-seqids?,INSDSeq_feature-table?,INSDSeq_sequence) >

<!--INSDSeq_length
-->
<!ELEMENT INSDSeq_length  (#PCDATA) >

<!--INSDSeq_moltype
Admissible values: DNA, RNA, AA
-->
<!ELEMENT INSDSeq_moltype  (#PCDATA) >

<!--INSDSeq_division
Indication that a sequence is related to a patent application. Must be populated with the
value PAT.
-->
<!ELEMENT INSDSeq_division  (#PCDATA) >

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

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

```
-->
<!ELEMENT INSDSeq_feature-table  (INSDFeature+) >

<!--INSDSeq_sequence
The residues of the sequence. The sequence must not contain numbers, punctuation or
whitespace characters.
-->
<!ELEMENT INSDSeq_sequence  (#PCDATA) >

<!--INSDSeqid
Intended for the use of Patent Offices in data exchange only.

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

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

Example:
pat|WO|2013999999|A1|123456

This represents the 123456th sequence from WO patent publication No. 2013999999 (A1)
-->
<!ELEMENT INSDSeqid  (#PCDATA) >

<!--INSDFeature
Description of one feature.
-->
<!ELEMENT INSDFeature  (INSDFeature_key,INSDFeature_location,INSDFeature_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.
-->
<!ELEMENT INSDQualifier  (INSDQualifier_name,INSDQualifier_value?) >

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

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

[Annex III to ST.26 follows]

ST.26 - ANNEX III

SEQUENCE LISTING SPECIMEN (XML file)

Final Draft

Proposal presented by the SEQL Task Force for consideration and adoption at the CWS/4

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ST26SequenceListing PUBLIC "-//WIPO//DTD SEQUENCE LISTING 1.0//EN""
ST26SequenceListing_V1_0.dtd">
<ST26SequenceListing dtdVersion="V1_0" fileName="AnnexIII_Sequence_Listing_Specimen.xml"
softwareName="SEQL-software-name" softwareVersion="1.0" productionDate="2013-12-17">
     <ApplicationIdentification>
          <IPOfficeCode>IB</IPOfficeCode>
          <ApplicationNumberText>PCT/IB2015/099999</ApplicationNumberText>
          <FilingDate>2015-01-31</FilingDate>
     </ApplicationIdentification>
     <ApplicantFileReference>AB123</ApplicantFileReference>
     <EarliestPriorityApplicationIdentification>
          <IPOfficeCode>IB</IPOfficeCode>
          <ApplicationNumberText>PCT/IB2014/111111</ApplicationNumberText>
          <FilingDate>2014-01-30</FilingDate>
     </EarliestPriorityApplicationIdentification>
     <ApplicantName languageCode="JA">出願製薬株式会社</ApplicantName>
     <ApplicantNameLatin>Shutsugan Pharmaceuticals Kabushiki Kaisha</ApplicantNameLatin>
     <InventorName languageCode="JA">特許 太郎</InventorName>
     <InventorNameLatin>Taro Tokkyo</InventorNameLatin>
     <InventionTitle languageCode="JA">efgタンパク質のためのマウスabcd-1遺伝子</InventionTitle>
     <InventionTitle languageCode="EN">Mus musculus abcd-1 gene for efg protein
     </InventionTitle>
     <SequenceTotalQuantity>11</SequenceTotalQuantity>
     <SequenceData sequenceIDNumber="1">
          <INSDSeq>
               <INSDSeq_length>133</INSDSeq_length>
               <INSDSeq_moltype>DNA</INSDSeq_moltype>
               <INSDSeq_division>PAT</INSDSeq_division>
               <INSDSeq_feature-table>
                    <INSDFeature>
                         <INSDFeature_key>source</INSDFeature_key>
                         <INSDFeature_location>1..133</INSDFeature_location>
                         <INSDFeature_quals>
                              <INSDQualifier>
                                   <INSDQualifier_name>organism</INSDQualifier_name>
                                   <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
                              </INSDQualifier>
                              <INSDQualifier>
                                   <INSDQualifier_name>mol_type</INSDQualifier_name>
                                   <INSDQualifier_value>genomic DNA</INSDQualifier_value>
                              </INSDQualifier>
                         </INSDFeature_quals>
                    </INSDFeature>
               </INSDSeq_feature-table>
               <INSDSeq_sequence>
               atgaaattaaaacataaaarggatgataaaatgagatttgatataaaaaggttttagagttagcagagaaggattttgaga
               cggcatggagagagacaagggcattaataaaggataaacatattgacaata</INSDSeq_sequence>
          </INSDSeq>
     </SequenceData>
     <SequenceData sequenceIDNumber="2">
          <INSDSeq>
               <INSDSeq_length>29</INSDSeq_length>
               <INSDSeq_moltype>AA</INSDSeq_moltype>
               <INSDSeq_division>PAT</INSDSeq_division>
               <INSDSeq_feature-table>
```

```
                <INSDFeature>
                    <INSDFeature_key>SOURCE</INSDFeature_key>
                    <INSDFeature_location>1..29</INSDFeature_location>
                    <INSDFeature_quals>
                        <INSDQualifier>
                            <INSDQualifier_name>ORGANISM</INSDQualifier_name>
                            <INSDQualifier_value>synthetic construct</INSDQualifier_value>
                        </INSDQualifier>
                        <INSDQualifier>
                            <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
                            <INSDQualifier_value>protein</INSDQualifier_value>
                        </INSDQualifier>
                        <INSDQualifier>
                            <INSDQualifier_name>NOTE</INSDQualifier_name>
                            <INSDQualifier_value>Synthetic peptide antigen fragment
                            </INSDQualifier_value>
                        </INSDQualifier>
                    </INSDFeature_quals>
                </INSDFeature>
            </INSDSeq_feature-table>
            <INSDSeq_sequence>GSLSDVRKDVEKRIDKALEAFKNKMDKEK</INSDSeq_sequence>
        </INSDSeq>
    </SequenceData>
    <SequenceData sequenceIDNumber="3">
        <INSDSeq>
            <INSDSeq_length>62</INSDSeq_length>
            <INSDSeq_moltype>DNA</INSDSeq_moltype>
            <INSDSeq_division>PAT</INSDSeq_division>
            <INSDSeq_feature-table>
                <INSDFeature>
                    <INSDFeature_key>source</INSDFeature_key>
                    <INSDFeature_location>1..62</INSDFeature_location>
                    <INSDFeature_quals>
                        <INSDQualifier>
                            <INSDQualifier_name>organism</INSDQualifier_name>
                            <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
                        </INSDQualifier>
                        <INSDQualifier>
                            <INSDQualifier_name>mol_type</INSDQualifier_name>
                            <INSDQualifier_value>genomic DNA</INSDQualifier_value>
                        </INSDQualifier>
                    </INSDFeature_quals>
                </INSDFeature>
                <INSDFeature>
                    <INSDFeature_key>CDS</INSDFeature_key>
                    <INSDFeature_location>3..62</INSDFeature_location>
                    <INSDFeature_quals>
                        <INSDQualifier>
                            <INSDQualifier_name>translation</INSDQualifier_name>
                            <INSDQualifier_value>MLAPDCPFDPTRIYSSSLC</INSDQualifier_value>
                        </INSDQualifier>
                        <INSDQualifier>
                            <INSDQualifier_name>protein_id</INSDQualifier_name>
                            <INSDQualifier_value>4</INSDQualifier_value>
                        </INSDQualifier>
                    </INSDFeature_quals>
                </INSDFeature>
            </INSDSeq_feature-table>
            <INSDSeq_sequence>tgatgctcgcacctgactgtcccttcgaccccacacgcatttatagctccagcctgtgctag
            </INSDSeq_sequence>
        </INSDSeq>
    </SequenceData>
    <SequenceData sequenceIDNumber="4">
        <INSDSeq>
            <INSDSeq_length>19</INSDSeq_length>
            <INSDSeq_moltype>AA</INSDSeq_moltype>
            <INSDSeq_division>PAT</INSDSeq_division>
```

```
            <INSDSeq_feature-table>
                  <INSDFeature>
                        <INSDFeature_key>SOURCE</INSDFeature_key>
                        <INSDFeature_location>1..19</INSDFeature_location>
                        <INSDFeature_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>
            <INSDSeq_sequence>MLAPDCPFDPTRIYSSSLC</INSDSeq_sequence>
      </INSDSeq>
</SequenceData>
<SequenceData sequenceIDNumber="5">
      <INSDSeq>
            <INSDSeq_length>133</INSDSeq_length>
            <INSDSeq_moltype>DNA</INSDSeq_moltype>
            <INSDSeq_division>PAT</INSDSeq_division>
            <INSDSeq_feature-table>
                  <INSDFeature>
                        <INSDFeature_key>source</INSDFeature_key>
                        <INSDFeature_location>1..133</INSDFeature_location>
                        <INSDFeature_quals>
                              <INSDQualifier>
                                    <INSDQualifier_name>organism</INSDQualifier_name>
                                    <INSDQualifier_value>Solanum lycopersicum</INSDQualifier_value>
                              </INSDQualifier>
                              <INSDQualifier>
                                    <INSDQualifier_name>mol_type</INSDQualifier_name>
                                    <INSDQualifier_value>genomic DNA</INSDQualifier_value>
                              </INSDQualifier>
                              <INSDQualifier>
                                    <INSDQualifier_name>note</INSDQualifier_name>
                                    <INSDQualifier_value>common name: tomato</INSDQualifier_value>
                              </INSDQualifier>
                        </INSDFeature_quals>
                  </INSDFeature>
                  <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>
                  <INSDFeature>
                        <INSDFeature_key>modified_base</INSDFeature_key>
                        <INSDFeature_location>22</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>
```

```
                    <INSDFeature>
                         <INSDFeature_key>variation</INSDFeature_key>
                         <INSDFeature_location>60</INSDFeature_location>
                         <INSDFeature_quals>
                              <INSDQualifier>
                                   <INSDQualifier_name>replace</INSDQualifier_name>
                                   <INSDQualifier_value>c</INSDQualifier_value>
                              </INSDQualifier>
                         </INSDFeature_quals>
                    </INSDFeature>
               </INSDSeq_feature-table>
               <INSDSeq_sequence>
               atgaaattaaaacanaaaaggnatgataaaatgagatttgatataaaaaaggttttagagttagcagagaaggattttgaga
               cggcatggagagagacaagggcattaataaaggataaacatattgacaata</INSDSeq_sequence>
          </INSDSeq>
     </SequenceData>
     <SequenceData sequenceIDNumber="6">
          <INSDSeq>
               <INSDSeq_length>29</INSDSeq_length>
               <INSDSeq_moltype>AA</INSDSeq_moltype>
               <INSDSeq_division>PAT</INSDSeq_division>
               <INSDSeq_feature-table>
                    <INSDFeature>
                         <INSDFeature_key>SOURCE</INSDFeature_key>
                         <INSDFeature_location>1..29</INSDFeature_location>
                         <INSDFeature_quals>
                              <INSDQualifier>
                                   <INSDQualifier_name>ORGANISM</INSDQualifier_name>
                                   <INSDQualifier_value>synthetic construct</INSDQualifier_value>
                              </INSDQualifier>
                              <INSDQualifier>
                                   <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
                                   <INSDQualifier_value>protein</INSDQualifier_value>
                              </INSDQualifier>
                              <INSDQualifier>
                                   <INSDQualifier_name>NOTE</INSDQualifier_name>
                                   <INSDQualifier_value>Synthetic peptide antigen fragment
                                   </INSDQualifier_value>
                              </INSDQualifier>
                         </INSDFeature_quals>
                    </INSDFeature>
                    <INSDFeature>
                         <INSDFeature_key>MOD_RES</INSDFeature_key>
                         <INSDFeature_location>3</INSDFeature_location>
                         <INSDFeature_quals>
                              <INSDQualifier>
                                   <INSDQualifier_name>NOTE</INSDQualifier_name>
                                   <INSDQualifier_value>N-acetylalanine</INSDQualifier_value>
                              </INSDQualifier>
                         </INSDFeature_quals>
                    </INSDFeature>
                    <INSDFeature>
                         <INSDFeature_key>SITE</INSDFeature_key>
                         <INSDFeature_location>7</INSDFeature_location>
                         <INSDFeature_quals>
                              <INSDQualifier>
                                   <INSDQualifier_name>NOTE</INSDQualifier_name>
                                   <INSDQualifier_value>Orn</INSDQualifier_value>
                              </INSDQualifier>
                         </INSDFeature_quals>
                    </INSDFeature>
                    <INSDFeature>
                         <INSDFeature_key>SITE</INSDFeature_key>
                         <INSDFeature_location>13</INSDFeature_location>
                         <INSDFeature_quals>
                              <INSDQualifier>
                                   <INSDQualifier_name>NOTE</INSDQualifier_name>
```

```
                                <INSDQualifier_value>D-Arginine</INSDQualifier_value>
                            </INSDQualifier>
                        </INSDFeature_quals>
                    </INSDFeature>
                    <INSDFeature>
                        <INSDFeature_key>UNSURE</INSDFeature_key>
                        <INSDFeature_location>15</INSDFeature_location>
                        <INSDFeature_quals>
                            <INSDQualifier>
                                <INSDQualifier_name>NOTE</INSDQualifier_name>
                                <INSDQualifier_value>A or V</INSDQualifier_value>
                            </INSDQualifier>
                        </INSDFeature_quals>
                    </INSDFeature>
                    <INSDFeature>
                        <INSDFeature_key>VARIANT</INSDFeature_key>
                        <INSDFeature_location>20</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>SITE</INSDFeature_key>
                        <INSDFeature_location>22</INSDFeature_location>
                        <INSDFeature_quals>
                            <INSDQualifier>
                                <INSDQualifier_name>NOTE</INSDQualifier_name>
                                <INSDQualifier_value>Homoserine</INSDQualifier_value>
                            </INSDQualifier>
                        </INSDFeature_quals>
                    </INSDFeature>
                </INSDSeq_feature-table>
                <INSDSeq_sequence>GSASDVXKDVEKRIXKALEXFSNKMDKSK</INSDSeq_sequence>
            </INSDSeq>
        </SequenceData>
        <SequenceData sequenceIDNumber="7">
            <INSDSeq>
                <INSDSeq_length/>
                <INSDSeq_moltype/>
                <INSDSeq_division/>
                <INSDSeq_sequence>000</INSDSeq_sequence>
            </INSDSeq>
        </SequenceData>
        <SequenceData sequenceIDNumber="8">
            <INSDSeq>
                <INSDSeq_length>74</INSDSeq_length>
                <INSDSeq_moltype>RNA</INSDSeq_moltype>
                <INSDSeq_division>PAT</INSDSeq_division>
                <INSDSeq_feature-table>
                    <INSDFeature>
                        <INSDFeature_key>source</INSDFeature_key>
                        <INSDFeature_location>1..74</INSDFeature_location>
                        <INSDFeature_quals>
                            <INSDQualifier>
                                <INSDQualifier_name>organism</INSDQualifier_name>
                                <INSDQualifier_value>Dengue virus 2</INSDQualifier_value>
                            </INSDQualifier>
                            <INSDQualifier>
                                <INSDQualifier_name>mol_type</INSDQualifier_name>
                                <INSDQualifier_value>genomic RNA</INSDQualifier_value>
                            </INSDQualifier>
                        </INSDFeature_quals>
                    </INSDFeature>
```

```
            </INSDSeq_feature-table>
            <INSDSeq_sequence>
            atgaaattaaaacataaaagggatgataaaatgagatttgatataaaaaaggttttagagttagcagagaagga
            </INSDSeq_sequence>
      </INSDSeq>
 </SequenceData>
 <SequenceData sequenceIDNumber="9">
      <INSDSeq>
            <INSDSeq_length>120</INSDSeq_length>
            <INSDSeq_moltype>DNA</INSDSeq_moltype>
            <INSDSeq_division>PAT</INSDSeq_division>
            <INSDSeq_feature-table>
                  <INSDFeature>
                        <INSDFeature_key>source</INSDFeature_key>
                        <INSDFeature_location>1..120</INSDFeature_location>
                        <INSDFeature_quals>
                              <INSDQualifier>
                                    <INSDQualifier_name>organism</INSDQualifier_name>
                                    <INSDQualifier_value>synthetic construct</INSDQualifier_value>
                              </INSDQualifier>
                              <INSDQualifier>
                                    <INSDQualifier_name>mol_type</INSDQualifier_name>
                                    <INSDQualifier_value>other DNA</INSDQualifier_value>
                              </INSDQualifier>
                        </INSDFeature_quals>
                  </INSDFeature>
                  <INSDFeature>
                        <INSDFeature_key>misc_feature</INSDFeature_key>
                        <INSDFeature_location>1..60</INSDFeature_location>
                        <INSDFeature_quals>
                              <INSDQualifier>
                                    <INSDQualifier_name>note</INSDQualifier_name>
                                    <INSDQualifier_value>DNA fragment</INSDQualifier_value>
                              </INSDQualifier>
                        </INSDFeature_quals>
                  </INSDFeature>
                  <INSDFeature>
                        <INSDFeature_key>misc_feature</INSDFeature_key>
                        <INSDFeature_location>61..120</INSDFeature_location>
                        <INSDFeature_quals>
                              <INSDQualifier>
                                    <INSDQualifier_name>note</INSDQualifier_name>
                                    <INSDQualifier_value>RNA fragment</INSDQualifier_value>
                              </INSDQualifier>
                        </INSDFeature_quals>
                  </INSDFeature>
            </INSDSeq_feature-table>
            <INSDSeq_sequence>
            cgacccacgcgtccgaggaaccaaccatcacgtttgaggacttcgtgaaggaattggataatacccgtccctaccaaaatgg
            cgagcgccgactcattgctcctcgtaccgtcgagcggc</INSDSeq_sequence>
      </INSDSeq>
 </SequenceData>
 <SequenceData sequenceIDNumber="10">
      <INSDSeq>
            <INSDSeq_length>288</INSDSeq_length>
            <INSDSeq_moltype>DNA</INSDSeq_moltype>
           <INSDSeq_division>PAT</INSDSeq_division>
            <INSDSeq_feature-table>
                  <INSDFeature>
                        <INSDFeature_key>source</INSDFeature_key>
                        <INSDFeature_location>1..288</INSDFeature_location>
                        <INSDFeature_quals>
                              <INSDQualifier>
                                    <INSDQualifier_name>organism</INSDQualifier_name>
                                    <INSDQualifier_value>Candida albicans</INSDQualifier_value>
                              </INSDQualifier>
                              <INSDQualifier>
```

```
                                    <INSDQualifier_name>mol_type</INSDQualifier_name>
                                    <INSDQualifier_value>genomic DNA</INSDQualifier_value>
                               </INSDQualifier>
                          </INSDFeature_quals>
                     </INSDFeature>
                     <INSDFeature>
                          <INSDFeature_key>CDS</INSDFeature_key>
                          <INSDFeature_location>1..288</INSDFeature_location>
                          <INSDFeature_quals>
                          <INSDQualifier>
                                    <INSDQualifier_name>translation</INSDQualifier_name>
                                    <INSDQualifier_value>
                                    MNLTLHNVIQTDSRGEKFMKIPEIYIRGIHIKYLRIPDDIMGYAKEQSMINMENRNRYQKRRGTSS
                                    GGGGGGGGGSGDSRRFNNRQSHGHNYGRR</INSDQualifier_value>
                               </INSDQualifier>
                               <INSDQualifier>
                                    <INSDQualifier_name>transl_table</INSDQualifier_name>
                                    <INSDQualifier_value>12</INSDQualifier_value>
                               </INSDQualifier>
                               <INSDQualifier>
                                    <INSDQualifier_name>protein_id</INSDQualifier_name>
                                    <INSDQualifier_value>11</INSDQualifier_value>
                               </INSDQualifier>
                          </INSDFeature_quals>
                     </INSDFeature>
                </INSDSeq_feature-table>
                <INSDSeq_sequence>
                atgaatttaaccttacataatgttatacaaaccgattcccgaggtgagaaatttatgaaaattcccgaaatatatattcgtg
                gtatacatattaaatatttaagaattcctgatgatattatgggatatgcaaaagaacaaagtatgataaatatggaaaatag
                aaatcgataccaaaaaagaagaggtactagcagtggtggtggtggtggtggtggtggaagtggtgattcaagaaggttt
                aataatagacaactgcatggacataattatggacgtagatga</INSDSeq_sequence>
           </INSDSeq>
      </SequenceData>
      <SequenceData sequenceIDNumber="11">
           <INSDSeq>
                <INSDSeq_length>95</INSDSeq_length>
                <INSDSeq_moltype>AA</INSDSeq_moltype>
                <INSDSeq_division>PAT</INSDSeq_division>
                <INSDSeq_feature-table>
                     <INSDFeature>
                          <INSDFeature_key>SOURCE</INSDFeature_key>
                          <INSDFeature_location>1..95</INSDFeature_location>
                          <INSDFeature_quals>
                               <INSDQualifier>
                                    <INSDQualifier_name>ORGANISM</INSDQualifier_name>
                                    <INSDQualifier_value>Candida albicans</INSDQualifier_value>
                               </INSDQualifier>
                               <INSDQualifier>
                                    <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
                                    <INSDQualifier_value>protein</INSDQualifier_value>
                               </INSDQualifier>
                          </INSDFeature_quals>
                     </INSDFeature>
                </INSDSeq_feature-table>
                <INSDSeq_sequence>
                MNLTLHNVIQTDSRGEKFMKIPEIYIRGIHIKYLRIPDDIMGYAKEQSMINMENRNRYQKRRGTSSGGGGGGGGGSGDSRRF
                NNRQSHGHNYGRR</INSDSeq_sequence>
           </INSDSeq>
      </SequenceData>
</ST26SequenceListing>
```

[Annex IV to ST.26 follows]

ST.26 - ANNEX IV

CHARACTER SUBSET FROM THE UNICODE BASIC LATIN CODE TABLE

Final Draft

Proposal presented by the SEQL Task Force for consideration and adoption at the CWS/4

The ampersand character (0026) is only permitted as part of a predefined entity or as part of a numeric character reference (&#nnnn;). The quotation mark (0022), the apostrophe (0027), the less-than sign (003C), and the greater-than sign (003E) are not permitted and must be represented by their predefined entities.

Unicode code point	Character	Name
0020		SPACE
0021	!	EXCLAMATION MARK
0023	#	NUMBER SIGN
0024	$	DOLLAR SIGN
0025	%	PERCENT SIGN
0026	&	AMPERSAND
0028	(	LEFT PARENTHESIS
0029	)	RIGHT PARENTHESIS
002A	*	ASTERISK
002B	+	PLUS SIGN
002C	,	COMMA
002D	-	HYPHEN-MINUS
002E	.	FULL STOP
002F	/	SOLIDUS
0030	0	DIGIT ZERO
0031	1	DIGIT ONE
0032	2	DIGIT TWO
0033	3	DIGIT THREE
0034	4	DIGIT FOUR
0035	5	DIGIT FIVE
0036	6	DIGIT SIX
0037	7	DIGIT SEVEN
0038	8	DIGIT EIGHT
0039	9	DIGIT NINE
003A	:	COLON
003B	;	SEMICOLON
003D	=	EQUALS SIGN
003F	?	QUESTION MARK
0040	@	COMMERCIAL AT
0041	A	LATIN CAPITAL LETTER A
0042	B	LATIN CAPITAL LETTER B
0043	C	LATIN CAPITAL LETTER C
0044	D	LATIN CAPITAL LETTER D
0045	E	LATIN CAPITAL LETTER E
0046	F	LATIN CAPITAL LETTER F
0047	G	LATIN CAPITAL LETTER G
0048	H	LATIN CAPITAL LETTER H
0049	I	LATIN CAPITAL LETTER I
004A	J	LATIN CAPITAL LETTER J
004B	K	LATIN CAPITAL LETTER K
004C	L	LATIN CAPITAL LETTER L
004D	M	LATIN CAPITAL LETTER M
004E	N	LATIN CAPITAL LETTER N
004F	O	LATIN CAPITAL LETTER O
0050	P	LATIN CAPITAL LETTER P
0051	Q	LATIN CAPITAL LETTER Q
0052	R	LATIN CAPITAL LETTER R
0053	S	LATIN CAPITAL LETTER S
0054	T	LATIN CAPITAL LETTER T
0055	U	LATIN CAPITAL LETTER U

Unicode code point	Character	Name
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

[Annex V to ST.26 follows]

ST.26 - ANNEX V

ADDITIONAL DATA EXCHANGE REQUIREMENTS (FOR PATENT OFFICES ONLY)

Final Draft

Proposal presented by the SEQL Task Force for consideration and adoption at the CWS/4

In the context of data exchange with database providers (INSD members), the Patent Offices should populate for each sequence the element `INSDSeq_other-seqids` with one `INSDSeqid` containing a reference to the corresponding published patent and the sequence identification number in the following 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 as set forth in ST.3; document kind code is the code for the identification of different kinds of patent documents as set forth in ST.16; publication number is the publication number of the application or patent; and Sequence identification number is the number of the sequence in that application or patent.

Example:
pat|WO|2013999999|A1|123456

Which would be translated into a valid XML instance as:

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

Where “123456” is the 123456th sequence from the WO publication no. 2013999999 (A1).

[End of Annex II and of document]