WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

E

WIPO/WEBINAR/STANDARDS/2021/4
ORIGINAL: ENGLISH
DATE: MAY 28, 2021

Webinar: WIPO ST.26 Advanced Module

hosted by the International Bureau of WIPO

Virtual, May 19, 2021
13:00 – 14:30 (CEST)

RESPONSES TO QUESTIONS

prepared by the USPTO and the International Bureau of WIPO

Following are the responses to the questions raised at the Webinar.

Q1: If a sequence with modified bases or amino acids is imported in WIPO Sequence, will the user get a warning stating that any feature keys and/or qualifiers are missing, if applicable?

A1: WIPO Sequence will not detect that a residue is modified until the user adds a feature key and qualifier indicating the modification. For example, if a user wants to include an amino acid sequence with a single D-alanine in a sequence listing, the sequence that is "introduced" (typed-in, pasted in, or imported in) will simply have an "A" at the position of the D-alanine. Until the user adds a feature and qualifier indicating that the alanine is a D-alanine, WIPO sequence will simply treat the "A" as an L-alanine.

If a user enters a feature key into a sequence, such as "modified_base" but does not include the mandatory qualifier "mod_base", WIPO Sequence will list an error on validation.

If a user includes the variable "n" in a nucleotide sequence or the variable "X" in an amino acid sequence, and does not include an annotation that describes the value of the "n" or "X" residue, WIPO Sequence will not list an error or warning on validation. In ST.26, "n" and "X" have default values, so it is perfectly acceptable to have these residues in a sequence and not include a description. The user must ensure that all "n" and "X" residues are properly described, if required.

Q2: Does the WIPO Sequence Validator check all of the things that you described during your presentation?

A2: The validator integrated into the WIPO Sequence desktop authoring tool is able to check most of the requirements described in this presentation.

For example, the validator will list an error if a feature key is missing a mandatory qualifier, or if a qualifier with a defined list of value choices has an inappropriate value. The validator will perform several checks on CDS features, ensuring that the value in the translation qualifier matches the theoretical translation of the CDS feature, taking into account any "transl_except", "codon_start", or "transl_table" qualifiers. It will also ensure that the value of the protein identified in the "protein_id" qualifier matches the value of the "translation" qualifier. The validator will also check location formats for feature keys to ensure the format is compliant. However, it is ultimately the responsibility of the applicant to ensure that a sequence listing accurately and completely describes sequences disclosed in a patent application.

The WIPO Sequence validator cannot guarantee an error-free sequence listing because certain values cannot be checked in an automated way, and require human review. For example, if the user errs in the entry of a custom organism name, the validator will only generate a warning to the user to confirm that the entry is correct. By definition, custom organism names are not included in the comprehensive list of scientific organism names contained in WIPO Sequence and, accordingly, cannot be verified by the tool. During human review of the sequence listing for compliance with the Standard after submission to an IPO, the error in the custom organism name could be found.

Q3: How do you represent a modified amino acid residue if the modification is not listed in ST.26 (e.g. peptide synthesis and engineering)?

A3: Within a sequence, a modified amino acid should be represented with the corresponding unmodified amino acids whenever possible. Where a modified amino acid in a sequence cannot be represented by any symbol in Annex 1, Table 3, the amino acid must be represented by the symbol "X". The modified amino acid residue must then be annotated using either a "SITE" feature key or a "MOD_RES" feature key, both of which have a mandatory "NOTE" qualifier. The qualifier "MOD_RES" is used only for post-translational modifications. Regardless of the feature key used to describe the modified amino acid, the value for the

“NOTE” qualifier should be the complete unabbreviated name of the modified amino acid (see paragraph 30).

Q4: Is it preferred to represent a sequence (or two sequences depending on your perspective) having a single point deletion using the qualifier "replace" with no value or by using two separate sequences (one with the residue and one without) or both? If both, do you then need three sequences, one with the qualifier and two without?

A4: The number of sequences that are required to be included in the sequence listing will depend on how the variants are disclosed in the application. ST.26 paragraphs 93-95 should be consulted for the representation of variants, in general. If the two sequences are separately enumerated (see paragraph 93), then both must be included in the sequence listing. If only a single sequence is enumerated, and the variant with a point deletion is described in prose (see paragraph 94), then only one sequence must be included in the sequence listing. This primary sequence must be annotated with a “replace” qualifier to indicate the location of the deletion.

Note that it is recommended that the sequences of all variants which are important to the invention being claimed, be included separately in the sequence listing, even if inclusion is not required by the Standard.

Q5: When I enter the national phase as a designated Office for a PCT application, are you saying that the patent office might require the applicant to provide a translation of the qualifier into a language used by that Office?

A5: That is correct. An IPO might require an applicant to provide a translation of the sequence listing where the language dependent free text qualifier values are in the language used by that Office.

Q6: Must sequence listings always be provided in English? Can I have a sequence listing in German as the main language? Or can only English be the main language?

A6: The value for each language-dependent qualifier in a sequence listing must be entered in the same language. This language may be English or another language, such as German, in accordance with the language(s) accepted by the particular Authority for that purpose. The Standard also permits a sequence listing to include all of the language-dependent qualifier values in English and, additionally, in a non-English language. So, yes, it is possible to create a sequence listing where all of the language dependent free text qualifier values are only in German.

Q7: In this example of a hybrid DNA/RNA sequence (see slide-deck), why is it required to specify 'misc_feature' as DNA, if the mol_type is DNA?

A7: Where an application discloses a DNA/RNA hybrid sequence, the sequence must be included in the sequence listing with the molecule type “DNA”, and the value for the mandatory “mol_type” qualifier of the “source” feature key is “other DNA”. ST.26 paragraph 55 specifically requires that “[e]ach DNA and RNA segment of the combined DNA/RNA molecule must be further described with the feature key “misc_feature” and the qualifier “note, which indicates whether the segment is DNA or RNA”. Thus, in a DNA/RNA hybrid molecule, the DNA segments must be indicated with a “misc_feature” despite the fact that the molecule type is DNA. This will ensure a complete and accurate description of every segment in the molecule.

Q8: In the hybrid DNA/RNA sequence example (feature location of 1-6 and 7-26 and 27-32), is it required that they appear in the generated XML sequence listing in that order? Or could the three feature keys be in any order, so long as it accounts for all of the locations?

A8: There is no requirement in ST.26 that feature keys appear in any particular order. For a DNA/RNA hybrid molecule, each DNA and RNA segment must be accounted for with a misc_feature feature key (i.e. every residue must be included in a misc_feature feature key), but no particular order is required. However, for clarity and ease of understanding, we do recommend including features in sequential order, where possible.

Q9: When importing a ST.25 sequence listing, will the WIPO Sequence desktop automatically bulk edit all the sequences with a 'u' residue?

A9: When importing an RNA sequence from an ST.25 sequence listing, all “u” symbols will be replaced by “t” symbols. This change will be noted in the import report.

When importing a DNA sequence from an ST.25 sequence listing, any “u” symbols will be maintained and will NOT be converted to “t”. WIPO Sequence cannot determine if a “u” in a DNA sequence is a modified residue (a uracil nucleobase on a DNA backbone) or an RNA segment of a DNA/RNA hybrid molecule. Therefore, maintaining the “u” symbols in DNA sequences will result in errors when the project is validated. These errors will require the user to manually change the “u” symbols to “t” symbols and prompt the inclusion of necessary feature keys and qualifiers that explain the presence of uracil in a DNA sequence.

Q10: If I want to provide the language dependent text in two languages using the WIPO Sequence desktop tool, then is it necessary for one of the language to be English?

A10: Yes, one of the languages must be English.

Q11: How could you annotate a sequence where phosphorothioate internucleotide linkage replaces the natural phosphate internucleotide linkage?

A11: A nucleotide sequence with one or more phosphorothioate internucleotide linkages can be annotated as a nucleotide analog under ST.26. The residues linked with phosphorothioate bonds must be identified using a “modified_base” feature key, a “mod_base” qualifier with the value “OTHER”, and a “note” qualifier that describes the phosphorothioate bond.

A single phosphorothioate bond between two adjacent residues can be described in a “modified_base” feature with the location format x^y, where x and y are the positions of the residues linked by the phosphorothioate bond. A region of contiguous residues that are all linked by phosphorothioate bonds can be described in a “modified_base” feature with the location format x..y, where x is the first residue in the region and y is the last residue in the region.

Q12: If you have a branched sequence where one of the branches is smaller than 4 AAs, what is the preferred method to disclose this sequence in a sequence listing? Should the MOD_RES feature be used, with the qualifier being the less than 4 AAs?

A12: If a branched amino acid sequence has a branch with <4 specifically defined amino acids, that particular branch cannot be included in the sequence listing as a separate sequence. However, the sequence listing must include any linear region of the branched sequence that contains >4 specifically defined amino acids. Where the branch included in the sequence listing links to a branch with <4 specifically defined amino acids, the amino acid involved in the linkage must be described using the feature key “SITE” and qualifier “NOTE”. The “NOTE” qualifier indicates that the residue is “bonded to a peptide of the sequence <insert sequence of <4 amino acid branch>”. The feature key “MOD_RES” is used only for post-translation modifications. If the linkage to the branch with <4 specifically defined amino acids is a result of a post-translation modification, then the feature key “MOD_RES” must be used with qualifier “NOTE” for the amino acid that links the branch with <4 specifically defined amino acids.

Q13: In your variant example (see slide 112), if you include the three specific variants in the sequence listing, should any feature or qualifier be included to specify that the variants are related to the primary generic sequence?

A13: While not required, applicants are always encouraged to include as much sequence annotation as possible in their sequence listing. An easy way to identify a specific variant as related to a primary generic sequence is to include a “note/NOTE” qualifier in the “source” or “SOURCE” feature that describes the relationship of the sequences.

Q14: For the sequence variant example (see slide 113), what if the sequence encompassing the variants covers more than a 1000 variants? This would be very time consuming to detail all of them.

A14: The main reason for ST.26 is to ensure that sequence data complies with INSDC (International Nucleotide Sequence Database Collaboration) requirements so that it can be included in public databases, and be made available to the public in as complete a form as possible. Paragraph 95(b), as explained in slide 113, requires the “most encompassing embodiment” to be included in the sequence listing when a variant “contains variations at multiple distinct locations and the occurrence of those variations are interdependent”. It is highly recommended that specific variants essential to the disclosure or claims of the invention be included as separate sequences, but it is not required.

If a patent application discloses a sequence with 1000 or more variants, the manner in which the variants are disclosed will dictate how they must be included in a sequence listing. ST.26 paragraphs 93-95 will control. It is the applicant’s responsibility to include disclosed sequence data in their sequence listing in compliance with the Standard, and to be as accurate and complete as possible with the annotations. This ensures that the full scope of applicant’s invention, as it pertains to biological sequences, is made readily available to IP Offices and to the public.

Q15: Why do *Ile* and *Leu* count as being similar enough for the “J” category, whereas *Ser* and *Thr* are not?

A15: ST.26 was designed to ensure sequence data was in a format compliant with INSDC (International Nucleotide Sequence Database Collaboration) requirements. Therefore, the amino acid symbols permitted in ST.26 are the amino acid symbols defined by the INSDC. The INSDC does not have a symbol for “Ser or Thr”, but does have a symbol (J) for “Leucine or Isoleucine”.

Q16: For residues which are repeated consecutively in a sequence (i.e., 5 glycines), do all the residues have to be listed in the sequence? Or is it ok to just annotate a residue being repeated a certain number of times?

Q16: ST.26 treats multiple residues represented by a shorthand formula as if each residue was separately enumerated (See ST.26 paragraph 3(c)(ii), and Annex VI, Introduction and Example 3(c)-2). Therefore, a disclosure with a shorthand formula such as “His6Gln” will be treated as if the disclosure was written out in long form: “His His His His His His Gln”. Accordingly, the sequence must be represented in the sequence listing as “HHHHHHQ”.

A sequence with repeating residues, such as a string of 5 glycines, should not be included in the sequence listing with a single instance of the repeating residue and annotation stating that the residue is repeated ‘x’ number of times. All residues must be represented individually in the sequence listing in this circumstance.

Note, however, that feature keys and qualifiers can be used to annotate variants of a primary sequence that differ in the number of times that a subsequence is repeated. For example, a sequence with a repeated region (e.g., can be annotated to describe variants with alternative numbers of repeats (e.g., using feature keys REPEAT, rpt_type, rpt_unit_range, or rpt_unit_seq) to describe variants with alternative numbers of repeats.

[End of document]