EDIT - bug #9527

Consensus Sequence Details: Consensus Sequence 5' -> 3' limit characters to [aAcCgGTtUu\s]

03/12/2021 02:59 PM - Andreas Kohlbecker

Status:	New	**Start date:**	
Priority:	New	**Due date:**	
Assignee:	Katja Luther	**% Done:**	0%
Category:	taxeditor	**Estimated time:**	0:00 hour
Target version:	Unassigned CDM tickets		
Severity:	normal	**Found in Version:**	

Description

for data to be entered or modified into the "Consensus Sequence 5' -> 3'" text-area the allowed characters, to be typed or pasted must be limited to those that are being used as code for the nucelosides of DNA, it might be a good idea though to also allow uracil which replaces thymin in RNA.

Also whitespace must be allowed.

Depending on how the consensus sequence is used, the consensus sequence calculates the most frequently appearing nucleotide for every position or it shows which residues are conserved and which residues are variable. Consider the following example DNA sequence: A[CT]N{A}YR. In this notation, A means that an A is always found in that position; [CT] stands for either C or T; N stands for any base; and {A} means any base except A. Y represents any pyrimidine, and R indicates any purine. (see https://en.wikipedia.org/wiki/Consensus_sequence)

Therefore we also need to allow different kind of brackets, Y and R. Maybe there are other characters used in consensus sequences.

regex for validation of DNA and RNA sequences:

```
^[aAcCgGTtUuRrNnYy\s\{\}\[\]].*$
```

History

#1 - 03/12/2021 03:32 PM - Andreas Müller

Is there a reason for this requirment or is it only because we expect sequences to be like this.
And why is it a TaxEditor ticket? Is it only a requirement for entering data in TaxEditor or is it a meant a general constraint?

The reason why I aske is that the problem with such constraints is that dirty data or strangely formatted data are then difficult to enter (e.g. during automated imports). Therefore it is always a trade off between correctnes and usability.
So the question is if there is a reason why we need this correctnes (e.g. because we have a viewer for sequences that requires it). Otherwise I would suggest to make it a soft validation rule (giving a hint that data is not correct but not forbid it)

#2 - 03/15/2021 10:48 AM - Andreas Kohlbecker

- Description updated

Andreas Müller wrote:

> Is there a reason for this requirment or is it only because we expect sequences to be like this.

It prevents from entering false data in the editor.

> And why is it a TaxEditor ticket? Is it only a requirement for entering data in TaxEditor or is it a meant a general constraint?
> The reason why I aske is that the problem with such constraints is that dirty data or strangely formatted data are then difficult to enter (e.g. during automated imports). Therefore it is always a trade off between correctnes and usability.

To avoid problems during the import etc this ticket specifically dedicated to the taxeditor. I adapted the the ticket description a bit to make that more clear.

#3 - 03/15/2021 12:44 PM - Katja Luther

- Description updated

#4 - 03/15/2021 12:58 PM - Katja Luther

there was a related ticket already, I add it for discussion informations: #5057

#6 - 03/15/2021 02:19 PM - Andreas Müller

The given regex is not valid anymore for the extended findings we have now on the usage of brackets and Y (and maybe others).

#7 - 03/15/2021 02:22 PM - Andreas Müller

With the given uncertainties on additional information like brackets I suggest to use only soft validation (warning but not forbidding). But finally the users which are familiar with possible formats (also dirty data) should decide.

#8 - 03/16/2021 01:51 PM - Andreas Kohlbecker

- Description updated

the regex is now complete according to the notation found on https://en.wikipedia.org/wiki/Consensus_sequence