Lecture 14 - Regular Expressions (regex)

Write regular expressions to match the following prompts, or describe/give an example of a match for the regular expression:

1. Species names: *Littorina saxatilis*, *Rhagoletis pomonella*, *Homo sapiens*

2. A positive marker for Huntington's disease, with more than 35 tandem repeats of `CAG`. Revise to allow repeats of `CAA`, which also encodes glutamine.

3. Latitude and longitude measurements of the format: `41 42'14.5" N, 86 14'01.6" W`

4. `[^\t]+\t[^\t]+\t[^\t]+\t[^\t]+`

5. `(GT)+(G{3}T{3})+`

6. A eukaryotic messanger RNA: an `AUG` start codon, 30-1000 bases of `A`,`U`,`G`,or `C`, and a 5-10 base poly-`A` tail

7. ATP/GTP-binding site motif A: `[AG].{4}GK[ST]`

8. Citations within the text of a paper, of the format `[24]`, `[2,73]`, `[5-7]`, etc.

9. `-?[0-9]{1,2}\.[0-9]+[ ,\t]+-?[0-9]{1,3}\.[0-9]+`

10. You and your collaborators recorded dates of data collection differently and you must match all of the following date formats:

`07/08/2016`

`7.25.16`

`August 5, 2016`

`Sept.8 '16`

`08-2-16`

11. Utilizing grep, print to standard out the `accession version numbers, species, sample information, and gene` in `R.mendax.1.fasta`. (Note, you are grabbing one continuous string.)

12. One additional metacharacter is `|`, which represents `or` (separates alternative match possibilities). Utilizing grep, print to standard out the open reading frames in `R.mendax.1.fasta`. (Start codon: ATG, Stop codons: TAA,TAG,TGA)