Regular Expressions Practice

Yassene Mohammed

Center for Proteomics and Metabolomics

Leiden University Medical Center

Eukaryotic mRNA sequence

- Write a regex pattern that matches Eukaryotic mRNA sequence

- Eukaryotic mRNA sequence
  - An AUG start codon at the beginning of sequence
  - Followed by 30 and 1000 bases (A, U, G or C)
  - Followed by a poly-A tail of 5 to 10 bases at the end of sequence

*: matches at least 0 times
+: matches at least 1 times
?: matches at most 1 times
{n}: matches exactly n times
{n,}: matches at least n times
{n,m}: matches between n and m times
Quantifier followed by ?: lazy

^: matches the start of a string
$: matches the end of a string
\b: matches a word boundary (empty string at either edge of a word)
\K: sets the given position in the regex as the new "start" of the match.

- (dot): any single character

[...] a character list
[^...] an inverted character list
\ escape operator
| or operator
(...) group operator (use \\n where n is number of the group)

In silico tryptic digestion

- Write a regex pattern to match the cleavage sites of trypsin
- Extend the previous pattern to match tryptic peptides

- Trypsin cleaves after arginine (R) or lysine (K) unless they are followed by proline (P)
- Use insulin sequence as an example

>sp|P01308|INS_HUMAN Insulin OS=Homo sapiens GN=INS PE=1 SV=1
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQV
GQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN