IDC306: Biocomputing
Assignment-9
Date: 14 Mar 2024

1. Rewrite to Homo Sapiens to H. sapiens or any similar names of organisms (general regex)

2. Rewrite '>CAA57801.1= GFP [Mus Musculus]' to '>CAA57801_M.Musculus'. Note: there could be any number of characters/space between >ID. and '[' symbol

3. Rewrite ">NM_001300425.1 Drosophila melanogaster Akt kinase (Akt), transcript variant E, mRNA" as ">NM_001300425_D.melanogaster_Akt_mRNA". Note: the >NM is fixed for any mRNA, the transcript variant may or may not be present i.e. the character after a variant is optional.

The general format is '>NM_DIGITS.ONE-TWO_DIGITS TAXON SPECIES LONG-NAME (SHORT_NAME), transcript variant OPTIONAL_TYPE, mRNA

LONG-NAME-multiple words, SHORT_NAME-one word OPTIONAL_TYPE-one single word

4. Find the coding region using regex (start codon: ATG/GTG; stop codon: TAA, TAG, TGA). Use any sequence from fasta_file.txt

5. Parse the below text and extract content within each HTML tag

```
<html>
  <body>
    <p>This is example text. </p>
  </body>
</html>
```

Challenge: Write a parser, if you have more than one entity <x>TEXT</x> within one block.

6. Parse **email.txt** given below to extract the email id of the sender (From), receiver (To), subject (line), content (lines) and store it dict() object.

RFC822 Message body
Received: from webmail.iisermohali.ac.in ([unix socket])
Received: from webmail.iisermohali.ac.in (localhost [127.0.0.1])
From: "GA" <ga@iisermohali.ac.in>
Date: Sun, 1 Mar 2020 19:16:32 +0530
To: "USer" <user@iisermohali.ac.in>
Subject: "Header"
Dear Sir,

Accumsan felis leo suspendisse vehicula orci purus vestibulum neque praesent imperdiet fusce sem a parturient penatibus dictum at suspendisse varius libero iaculis ligula. Adipiscing platea est bibendum vivamus nascetur torquent mus augue auctor vestibulum eleifend eu risus et eros ac vehicula nunc posuere. Nostra gravida at natoque diam euismod auctor parturient hac per in suspendisse suspendisse orci a gravida vestibulum a facilisis natoque hac ac scelerisque a cursus fames id himenaeos. Suspendisse mus parturient consectetur adipiscing convallis mauris et posuere arcu adipiscing magnis nisl natoque mollis dis bibendum vulputate nisi bibendum parturient gravida. Parturient mauris in felis dignissim id auctor mi in malesuada a torquent dui pulvinar sociosqu nam nisl curabitur elementum. Eu nam dictumst at in euismod praesent arcu penatibus ligula viverra cras a parturient scelerisque sagittis a egestas pharetra adipiscing suspendisse. Magna condimentum duis a vestibulum placerat a ac metus condimentum a mi felis ultrices dolor sem adipiscing turpis.
Thanking you,
Sincerely,
phone no: 999999
Indian Institute of Science Education and Research (IISER), Mohali.