Esame 05/07/2012

Andrea Passerini
passerini@disi.unitn.it

Informatica

Programma python

Scrivere una funzione `UTRConservationHistogram(conserved_regions, k)` che prenda in ingresso:

- un file `conserved_regions` con informazioni su sottosequenze conservate in sequenze UTR (Untranslated mRNA regions)
- un intero `k` che indica la lunghezza delle sottostringhe per gli istogrammi

calcoli, separatamente per 5UTR e 3UTR (che sono le UTR prima e dopo la sequenza codificante), gli istogrammi di frequenza delle sottostringhe di lunghezza *k* all'interno delle sottosequenze conservate e li stampi in uscita.

Esempio dati

```
$ cat conserved_regions
HGNC ucsc chr start end strand region.length region.sequence utr.type
SDF4 uc001adj.1 chr1 1159309 1159325 - 17 GAGGAACCGTGACTAGA 5UTR
SDF4 uc001adj.1 chr1 1159341 1159347 - 7 GTAGGTG 5UTR
DVL1 uc002quu.2 chr1 1275478 1275556 - 79  CTACCTCGGTTA... 5UTR
....
```

Esempio esecuzione

```
>>> import utility
>>> utility.UTRConservationHistogram('conserved_regions',3)
5UTR
ACC 7629
CTT 6624
ACA 7382
ACG 5317
ATC 4558
AAC 6992
ATG 4476
AGG 8391
CCT 7431
...     ...
3UTR
ACC 4008
ATG 6152
ACA 6186
ACG 2800
ATC 3665
AAC 5521
ATA 6558
AGG 4246
CCT 4531
...     ...
```

Programma python: suggerimento

Si possono implementare 5 funzioni separate:

1. una che estragga dal file le sottosequenze conservate, dividendole in 5UTR e 3UTR (e.g. dizionario)
2. una che prenda in ingresso il dizionario prodotto e separatamente per le sequenze in 5UTR e 3UTR chiami:
   - una funziona che calcola l'istogramma delle frequenze di sottostringhe di lunghezza k
   - una funziona che stampa l'istogramma
3. una che realizzi il programma richiesto usando le funzioni di cui sopra

Esercizi da linea di comando

- Contare quante sequenze nella directory `fasta` contengono almeno uno dei seguenti motivi di fosforilazione ("+" va letto come "seguito da"):
  - tirosina (Y) + metionina (M) + qualunque a-a + metionina (M)
  - tirosina (Y) + due qualunque a-a + prolina (P)
  - tirosina (Y) + acido glutamico (E) + qualunque a-a + valina (V) oppure isoleucina (I)
- Risposta: `26`

Esercizi da linea di comando

- L'ultimo carattere dell'identificatore dei file fasta indica quale catena della proteina di riferimento e' codificata nel file stesso. Sfruttando questo fatto, calcolare il numero di catene C ed il numero di catene D presenti nella directory `fasta`.
- Risposta:

```
5 C
6 D
```

- Suggerimento: `cut` puo' anche stampare caratteri in specifiche posizioni (e.g. il terzo carattere, vedere il man)