Esame 11/09/2014

Andrea Passerini
passerini@disi.unitn.it

Informatica

Programma python

Scrivere una funzione `findPerfectMatches(filename,pattern)` che:

- prenda in ingresso un nome di file `filename` con un elenco di sequenze di rna, e una stringa di rna (e.g. microRNA)
- stampi per ciascuna sequenza, il nome e l'elenco dei punti di match perfetto con la stringa (solo per le sequenze con almeno un match)

Esempio esecuzione

```
>>> import utility
>>> utility.findPerfectMatches('utr.txt','acgaatt')
ENSG00000050344 [1186]
ENSG00000109929 [204, 373, 3336]
ENSG00000155850 [2162, 5387]
ENSG00000073756 [1152]
ENSG00000175445 [693]
ENSG00000159167 [781]
ENSG00000138061 [1229]
ENSG00000152268 [1362]
ENSG00000197121 [3024]
ENSG00000115159 [1111]
ENSG00000169908 [1695]
ENSG00000150938 [1751]
ENSG00000179314 [2782]
ENSG00000075223 [1743]
```

Programma python: suggerimento

Si possono implementare 5 funzioni separate:

1. una che legga il file `filename` e restituisca un dizionario di coppie nome,sequenza
2. una che prenda in ingresso una stringa e ne restituisca il suo complementare (notare che le sequenze hanno la 't' e non la 'u' anche se rappresentano rna quindi anche la stringa in ingresso sarà fatta con 't')
3. una che prenda in ingresso due sequenze, e restituisca l'elenco dei punti di match della seconda sequenza sulla prima (guardare bene l'help della funzione `find`)
4. una che prenda in ingresso il dizionario di sequenze e la stringa convertita, e per ciascuna sequenza calcoli i punti di match chiamando la funzione precedente e se ci sono, stampi nome della sequenza ed elenco dei match
5. una che realizzi il programma richiesto usando le funzioni di cui sopra

Shell: esercizio #1

Dato il file `all.fasta`, calcolare quante sequenze contengono:

1. Una arginina (R), seguita da una glicina (G), seguita da acido aspartico (D).
2. Due proline (P), seguite da un amino acido qualunque, seguito da una leucina (L), seguita da una isoleucina (I).
3. Almeno uno dei due precedenti motivi.
4. Il primo motivo seguito (anche a distanza) dal secondo.

Soluzione

(1) 255, (2) 17, (3) 270, (4) 2.

Shell: esercizio #2

Dato il file `all.fasta`, contare quante proteine appartengono a ciascuna localizzazione cellulare — indicata nell'header delle varie sequenze.

Esempio

La proteina citoplasmatica `Q07815` è scritta:

```
>Q07815 cytoplasmic
MDGSGEQPRGGVSSRIEQGEWGGRHPSWPWTRCLRMRPPRS
```

Soluzione

510 mitochondrial; 837 nuclear; 843 extracellular; 1411 cytoplasmic

Modalita' di esecuzione e consegna

1. Avviare la macchina in modalita' `ESAME`
2. Autenticarsi con nome utente `sci-esame` e password fornita dal docente
3. Il testo del compito ed i file necessari si trovano in una cartella `Testo` sul Desktop
4. Realizzare il programma python come file `utility.py` e scrivere gli esercizi da linea di comando in un file di testo `linea_di_comando.txt`
5. Creare sul Desktop una cartella con *nome_cognome* e metterci i due file realizzati.
6. Eseguire il logout ma NON spegnere la macchina