Esame 12/01/2016

Andrea Passerini
passerini@disi.unitn.it

Informatica

Programma python

Scrivere un programma python che:

- prenda in ingresso:
  - un nome di file che contiene un elenco di siti di splicing D (donor) in formato fasta. Ogni sito contiene 15bp, 7 prima del sito e 8 dopo.
  - un valore di soglia
- stampi la frequenza di occorrenza di sottosequenze subito prima e subito dopo il sito, di lunghezza minima due e massima sette (prima del sito) e otto (dopo), limitatamente alle sottosequenze con frequenza pari almeno alla soglia.

Esempio ingresso

```
>HUMGLUT4B_2257
CTCCGAAGTAGGATT
>HUMGLUT4B_3651
TCAGAAGGTGAGGGC
>HUMGLUT4B_3935
TTGGAAGGTTCGCAG
>HUMGLUT4B_4152
TACTCAGGTACTCAC
>HUMGLUT4B_4379
CGCCCAGGTGACCGG
>HUMGLUT4B_4669
AGAAAGAGTAAGCTC
...
```

Esempio sottosequenze

Nota

Data una sequenza `CTCCGAAGTAGGATT` le sottosequenze sono:

- Prima del sito (quindi, andando a ritroso):

  `AA,GAA,CGAA,CCGAA,TCCGAA,CTCCGAA`

- Dopo il sito:

  `GT,GTA,GTAG,GTAGG,GTAGGA,GTAGGAT,GTAGGATT`

Esempio esecuzione

```
> python splice_patterns.py
Nome file: splice_donor.fasta
Soglia: 100
{'AAG': 208, 'AG': 583, 'GG': 107, 'GAG': 108,
 'TG': 135, 'CAG': 247}
{'GTA': 568, 'GT': 1116, 'GTAAGT': 126, 'GTGAG': 388,
 'GTG': 491, 'GTGA': 399, 'GTGAGT': 204, 'GTAA': 379,
 'GTAAG': 300, 'GTGAGTG': 116}
```

Programma python: suggerimento

Si possono implementare 5 funzioni separate:

1. Una che legga il file e restituisca un elenco di sequenze precedenti al sito, ed uno di sequenze successive.
2. Una che dato un elenco di sequenze precedenti al sito, restituisca un dizionario di occorrenze delle sottosequenze *terminali* di lunghezza da due a sette.
3. Una che dato un elenco di sequenze successive al sito, restituisca un dizionario di occorrenze delle sottosequenze *iniziali* di lunghezza da due a otto.
4. Una che dato un dizionario di conteggi ed una soglia, restituisca un dizionario con solo le coppie con valore maggiore o uguale alla soglia
5. una che realizzi il programma richiesto usando le funzioni di cui sopra

Shell: esercizio #1

Dati il file `sequences.fasta` ed i due seguenti motivi:

- una fenilalanina (F); due amminoacidi qualunque; una fenilalanina. Il motivo **deve** appare esattamente alla fine della sequenza.
- una arginina (R); una fenilalanina; un amminoacido qualunque tranne una prolina (P); una isoleucina (I) o una valina (V). Il motivo **non può** apparire alla fine della sequenza.

calcolare: (1) quante sequenze includono il primo motivo; (2) quante sequenze includono il secondo motivo; (3) quante sequenze includono almeno uno dei due motivi.

Risultato atteso

1. 4
2. 196
3. 200

Shell: esercizio #2

Dato il file `sequences.fasta`, e sfruttanto il formato delle intestazion (e.g. "`>WT1_HUMAN:Nucleus`"), stampare a schermo le quattro combinazioni di organismo e localizzazione subcellulare piu' frequenti.

Risultato atteso

```
636 HUMAN:Nucleus
203 HUMAN:Cytoplasm
191 HUMAN:Secretory
188 DROME:Nucleus
```

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