Esame 6/09/2018

Andrea Passerini
andrea.passerini@unitn.it

Informatica

Programma python

Scrivere un programma che prenda in ingresso un file di sequenze FASTA e un valore intero k e stampi per ogni proteina nel file:

- tutti i kmers di lunghezza k contenuti nella proteina, con le loro frequenze
- tutte le sequenze fatte dello stesso aminoacido contenute nella proteina, con le loro frequenze

Esempio esecuzione

```
> python repetitions.py
Inserire nome file fasta: sequences.fasta
Inserire lunghezza kmers: 3
16vpA
kmers
{'SRM': 1, 'RMP': 1, 'MPS': 1, 'PSP': 1, ...
repetitions
{'PP': 3, 'AA': 4, 'LL': 2, 'DD': 1, 'VV': 1, ...
1a3c_
...
```

Programma python: suggerimento

Si possono implementare 5 funzioni separate:

1. una che legga il file di sequenze e restituisca un dizionario da nome a sequenza
2. una che data una sequenza, ed un valore k, restituisca un dizionario con i kmer presenti nella sequenza e le loro frequenze
3. una che data una sequenza, restituisca un dizionario con le sottosequenze contigue dello stesso aminoacido presenti nella sequenza e le loro frequenze
4. una che dato un dizionario di sequenze e un valore k, calcoli e stampi per ogni sequenza i suoi kmers e le sue sottosequenze con ripetizioni dello stesso aminoacido
5. una (o un main) che realizzi il programma richiesto usando le funzioni di cui sopra

Esercizi da linea di comando

Stampare l'elenco dei cellular compartments (indicati da C) presenti nel file `gene-ontology.tab` con le rispettive frequenze, ordinati per frequenza decrescente.

Risultato atteso

```
2292 cytoplasm
1263 nucleus
 925 membrane
 587 endomembrane system
 580 mitochondrion
 367 endoplasmic reticulum
 270 vacuole
 270 plasma membrane
 240 chromosome
 186 ribosome
...
```

Esercizi da linea di comando

Calcolare quante tra le sequenze nella directory `fastas/` contengono una Alanina (A) seguita da uno di questi due pattern:

- una sequenza di tra due e cinque residui che non siano né Alanine (A), né Treonine (T) né Glicine (G), seguita da una Alanina (A) o una Glicina (G)
- una sequenza di tra uno e tre residui che non siano né Alanine (A), né Treonine (T) né Glicine (G), seguita da una Alanina (A)

Risultato atteso

```
89
```

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 `programma.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