Esame 19/07/2017

Andrea Passerini
passerini@disi.unitn.it

Informatica

Programma python

Scrivere un programma python che:

- prenda in ingresso un file contenente annotazioni di Gene Ontology per proteine, ed un file fasta con le sequenze aminoacidiche delle proteine.
- stampi per ciascun termine della Gene Ontology, la composizione aminoacidica media dell'insieme delle proteine annotate con quel termine

Esempio file Ingresso

File GO

```
...
!
!Generated: 2017-07-03 09:14
!GO-version: http://purl.obolibrary.org/obo/go/releases/2017-0
!
UniProtKB       A0A075B6H9      IGLV4-69    GO:0002250  ...
UniProtKB       A0A075B6H9      IGLV4-69    GO:0002377  ...
...
UniProtKB       A0A075B6I0      IGLV8-61    GO:0002250  ...
```

File Fasta

```
...
>sp|A0A075B6H9|LV469_HUMAN Immunoglobulin lambda variable 4-69
MAWTPLLFLTLLLHCTGSLSQLVLTQSPSASASLGASVKLTCTLSSGHSSYAIAWHQQQP
EKGPRYLMKLNSDGSHSKGDGIPDRFSGSSSGAERYLTISSLQSEDEADYYCQTWGTGI
>sp|A0A075B6I0|LV861_HUMAN Immunoglobulin lambda variable 8-61
MSVPTMAWMMLLLGLLAYGSGVDSQTVVTQEPSFSVSPGGTVTLTCGLSSGSVSTSYYPS
WYQQTPGQAPRTLIYSTNTRSSGVPDRFSGSILGNKAALTITGAQADDESDYYCVLYMGS
GI
...
```

Esempio esecuzione

```
> python go2sequences.py
nome file GO: go.txt
nome file Fasta: seqs.fasta
GO:0002250
{'M': 0.020912124582869854, 'A': 0.06963292547274749,
 'W': 0.022024471635150165, 'T': 0.0709677419354838 7,
 'P': 0.06384872080088988, 'L': 0.11434927697441602,
 'F': 0.027363737486095663, 'H': 0.011345939933259178,
 ...}
GO:0002377
{'M': 0.020202020202020204, 'A': 0.06788818416725394,
 'W': 0.021611463471928587, 'T': 0.06953253464881372,
 'P': 0.06741836974395114, 'L': 0.11627906976744186,
 'F': 0.026779422128259338, 'H': 0.010570824524312896
 ...}
...
```

Programma python: suggerimento

Si possono implementare 5 funzioni separate:

1. Una che legga il file di Gene Ontology e restituisca un dizionario che associa ad ogni termine GO la lista degli id delle proteine che lo contengono
2. Una che legga il file fasta e restituisca un dizionario id di proteina $\rightarrow$ sequenza aminoacidica
3. Una che data una sequenza aminoacidica restituisca un dizionario con la sua composizione media
4. Una che dati i dizionari gene ontology $\rightarrow$ nomi di proteine e id proteina $\rightarrow$ sequenza, per ogni termine concateni le proteine con quel termine, calcoli la composizione media usando la funzione precedente e la stampi
5. una che realizzi il programma richiesto usando le funzioni di cui sopra

Shell: esercizio #1

Calcolare quante sequenze nel file `seqs.fasta` contengono:

- Un acido aspartico (D) o glutammico (E) seguito da una glicina (G), seguita da due o tre aminoacidi qualunque, seguiti da una coppia alanina (A) e triptofano (W) o da una coppia cisteina (C) e acido glutammico (E).

Risultato atteso

```
9
```

Attenzione

Le sequenze proteiche possono occupare più di una riga

Shell: esercizio #2

Dato il file `go.txt`, estrarre i termini GO ordinati per frequenza di occorrenza

Risultato atteso

```
294 GO:0005515
241 GO:0005576
185 GO:0005886
150 GO:0016021
133 GO:0005829
 96 GO:0005634
 67 GO:0003823
 65 GO:0005737
...
```

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