Esame 16/06/2016

Andrea Passerini
passerini@disi.unitn.it

Informatica

Programma python

Scrivere un programma python che:

- prenda in ingresso un file che contiene in ogni riga l'associazione tra una proteina ed un dominio PFAM in essa identificato.
- stampi una lista di tutti i domini PFAM (con identificativo e nome), ordinati per frequenza di occorrenza, con l'aggiunta per ogni dominio del numero di proteine *distinte* che lo possiedono.

Esempio file

```
> cat output_pfam
```

Esempio esecuzione

```
> python pfam_statistics.py
Inserire nome file: output_pfam
Pfam Number Pfam Name        TotCount        ProtCount

PF00005.19      ABC_tran          35      32
PF00132.16      Hexapep 29        5
PF00534.12      Glycos_transf_1 21        20
PF00515.20      TPR_1   17        8
PF01370.13      Epimerase         16      16
PF03130.8       HEAT_PBS          15      6
PF04055.13      Radical_SAM       14      14
PF00502.11      Phycobilisome     13      12
PF00427.13      PBS_linker_poly 13        10
PF00271.23      Helicase_C        13      13
PF00528.14      BPD_transp_1      12      11
PF00004.21      AAA     12        12
PF00805.14      Pentapeptide      11      5
PF00535.18      Glycos_transf_2 11        11
PF00353.11      HemolysinCabind 11        3
PF07862.3       Nif11   10        10
PF01926.15      MMR_HSR1          10      9
...
```

Programma python: suggerimento

Si possono implementare 5 funzioni separate:

1. Una che legga il file e restituisca un dizionario da numero PFAM a nome PFAM, e per ogni proteina l'elenco dei suoi domini
2. Una che date le informazioni per proteina lette, restituisca un dizionario con il numero di occorrenze di ogni dominio PFAM
3. Una che date le informazioni per proteina lette, restituisca un dizionario con il numero di occorrenze di ogni dominio PFAM, contando una sola volta le occorrenze nella stessa proteina
4. una che stampi numero, nome e numero di occorrenze totali e per proteina di ciascun dominio, ordinando i domini per numero di occorrenze totali
5. una che realizzi il programma richiesto usando le funzioni di cui sopra

Shell: esercizio #1

Dato il file `output_pfam`, stampare a schermo i quattro domini di binding (contengono il testo "`_bind`") che occorrono più vicino all'inizio della sequenza.

Vanno stampati il nome del dominio, l'ID Pfam del dominio, e la posizione dell'istanza all'interno della sequenza, nell'ordine mostrato sotto.

Risultato atteso

```
1 158 PF03446.7 NAD_binding_2
1 294 PF04321.9 RmlD_sub_bind
1 32 PF01494.11 FAD_binding_3
1 90 PF00216.13 Bac_DNA_binding
```

Shell: esercizio #2

Dato il file `sequences.fasta`, contare quante sequenze contengono:

1. Una arginina (R) seguita da una lisina (K). Il motivo non deve trovarsi all'inizio della sequenza.
2. Due arginine seguite da un aminoacido che non sia nè una arginina, nè una lisina.
3. Nessuno dei due precedenti motivi.

usando il minor numero possibile di invocazioni a `grep`.

Risultato atteso

(1) 736 (2) 604 (3) 46.

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