Esame 30/06/2015

Andrea Passerini
passerini@disi.unitn.it

Informatica

Programma python

Scrivere un programma python che:

- prenda in ingresso un nome di file `filename` che contiene un elenco di occorrenze di domini su proteine.
- stampi per ciascun dominio la sua lunghezza media, ed il numero di proteine *distinte* in cui compare.

Esempio ingresso

UniProtKB Protein ID (Name)			Match Start Position		Match Stop Position			
1433E_HUMAN	4	237	0	255	37694	Pfam	IPR023410	14-3-3_domain
1433G_HUMAN	4	240	0	247	37694	Pfam	IPR023410	14-3-3_domain
1433Z_HUMAN	4	235	0	245	37694	Pfam	IPR023410	14-3-3_domain
4F2_HUMAN	239	320	3.70E-14		630	Alpha-amylase	Pfam	IPR006047
A1CF_HUMAN	58	124	4.50E-17		594	RRM_1	Pfam	IPR000504
A1CF_HUMA	138	199	0.00000016		594	RRM_1	Pfam	IPR000504
AATF_HUMAN	220	373	2.20E-38		560	AATF-Che1	Pfam	IPR025160
AATF_HUMAN	464	548	1.10E-27		560	TRAUB	Pfam	IPR012617

...

Esempio esecuzione

```
> python program.py
Name of file: domini.txt
domain                                avg_length     num_prots
Prot_Kinase_C-like_PE/DAG-bd          50.500000          1
CPL                                   148.000000         1
ATPase_F1/V1/A1_a/bsu_nucl-bd         225.000000         1
Dynein_heavy_dom                      725.000000         1
FHA_dom                               72.400000          5
Enolase_C                             289.000000         1
Syja_N                                287.500000         2
DBC1/CARP1_inactive_NUDIX_dom         126.000000         2
...
```

Programma python: suggerimento

Si possono implementare 5 funzioni separate:

1. una che legga il file e restituisca un dizionario con nome di dominio come chiave e come valore la lista delle occorrenze di quel dominio
2. una che prenda in ingresso il dizionario letto e restituisca un nuovo dizionario, con dominio come chiave e come valore la lunghezza media delle occorrenze di quel dominio
3. una che prenda in ingresso il dizionario letto e restituisca un nuovo dizionario, con dominio come chiave e come valore il numero di proteine *distinte* contenenti quel dominio.
4. una che prenda in ingresso i dizionari di lunghezze medie e numero di proteine e stampi per ogni dominio i valori corrispondenti
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:

- Un residuo qualunque seguito da una arginina (R) ed una lisina (K); **oppure** due arginine seguite da un residuo che non sia nè una arginina nè una lisina.
- Un acido aspartico (D), una alanina (A), un residuo qualunque, una glicina (G) ed un altro residuo qualunque.

calcolare quante sequence includono:

1. il primo motivo *seguito a distanza arbitraria* dal secondo.
2. il primo motivo ma *non* il secondo.

Risultato atteso

1. 64
2. 717

Shell: esercizio #2

Dato il file `domini.txt`, calcolare il numero di proteine distinte di lunghezza compresa tra 800 e 899 residui, estremi inclusi. (La lunghezza della proteina è annotata nella quinta colonna del file).

Risultato atteso

81

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