Esame 25/06/2013

Andrea Passerini
passerini@disi.unitn.it

Informatica

Programma python

Scrivere una funzione

`computeGeneRegulatoryElementCounts(genefile,clusterfile)`

che:

- prenda in ingresso un nome di file `genefile` con un elenco di geni e cluster ad essi associati, e un nome di file `clusterfile` con RBP e miRNA associati ad ogni cluster
- stampi per ogni gene il numero di RBP e di miRNA distinti che legano il suo mRNA

File dei geni

```
geneSymbol      cluster
AAA1    01
AAGAB   01
AAK1    01
AARSD1  01
AASDH   01
AASDHPPT        01
AASDHPPT        02
AASDHPPT        03
AASDHPPT        07
AASDHPPT        11
AASDHPPT        19
AASS    01
...
```

File dei clusters

```
cluster composition
01 ELAVL1
02 ELAVL1,EWSR1,FUS,MOV10,TAF15
03 ELAVL1,HNRNPD
04 PABPC1
05 ELAVL1,PUM2,hsa-miR-130a,hsa-miR-130b,hsa-miR-148a,hsa-miR-148b,hsa
06 ELAVL1,hsa-miR-15a,hsa-miR-15b,hsa-miR-16,hsa-miR-424
07 ELAVL1,MOV10,PUM2
08 ELAVL1,hsa-miR-106b,hsa-miR-148a,hsa-miR-17,hsa-miR-18a,hsa-miR-20a
09 ELAVL1,hsa-let-7a,hsa-let-7b,hsa-let-7c,hsa-let-7d,hsa-let-7e,hsa-l
10 ELAVL1,PUM1
11 ELAVL1,EWSR1,FUS,MOV10
12 ELAVL1,PUM2,hsa-miR-103,hsa-miR-107,hsa-miR-183,hsa-miR-221,hsa-miR
13 ELAVL1,U2AF2
14 ELAVL1,hsa-miR-103,hsa-miR-107,hsa-miR-15a,hsa-miR-15b,hsa-miR-16,h
15 DGCR8,ELAVL1
16 ELAVL1,PUM2,hsa-miR-101,hsa-miR-128,hsa-miR-27a,hsa-miR-27b
17 PARK7
18 CELF1,ELAVL1
19 ELAVL1,hsa-miR-1
20 ELAVL1,FUS,MOV10,RBFOX2
21 ELAVL1,hsa-miR-124
22 ELAVL1,MOV10,PUM2,hsa-miR-130a,hsa-miR-130b,hsa-miR-183,hsa-miR-25,
23 ELAVL1,EWSR1,FUS,MOV10,PUM2,QKI
24 ELAVL1,EWSR1,FUS,MOV10,PUM2,TNRC6B
```

Esempio esecuzione

```
>>> import utility
>>> utility.computeGeneRegulatoryElementCounts('genes.txt',
... 'clusters.txt')
gene RBPs miRNAs
RNF14 1 0
UBE2Q1 3 10
UBE2Q2 5 0
RNF10 2 0
RNF13 5 8
AK127179 1 0
LOC100506866 1 0
...
```

Programma python: suggerimento

Si possono implementare 4 funzioni separate:

1. una che legga il file dei geni e restituisca un dizionario con gene come chiave e lista di cluster id come valore
2. una che legga il file dei clusters e restituisca una dizionario con cluster id come chiave e lista di elementi regolatori (RBP e/o miRNA) come valore
3. una che prenda in ingresso i due dizionari e, per ogni gene con la sua lista di cluster id:
    1. crei un dizionario, e per ogni cluster id
    2. recuperi l'elenco di elementi regolatori
    3. aggiunga ciascuno al dizionario come chiave, mettendo come valore True se RBP, False se miRNA (le RBP si riconoscono perche' cominciano con una lettera maiuscola)
    4. conti il numero di RBP (True) e miRNA (False) trovati
    5. stampi nome del gene e conteggi trovati
4. una che realizzi il programma richiesto usando le funzioni di cui sopra

Esercizio Shell 1

Calcolare quanti file in `fasta/` contengono esattamente 93 e 132 **caratteri** (inclusi spazi, a capo, *etc.*), stampando a schermo **solo** quei casi.

Soluzione

5 catene contengono 93 caratteri.
8 catene contengono 132 caratteri.

Esercizio Shell 2

Contare quante catene in `fasta/` contengono uno dei seguenti motivi di cleavage:

- **PC7**: arginina (R), seguita da tre residui qualunque, seguiti da una arginina o una lisina (K), seguiti da una arginina, seguita da un residuo qualunque.
- **PACE**: arginina, seguita da un residuo qualunque, seguito da una arginina o una lisina, seguito da una arginina, seguita da un residuo qualunque.

Infine, quante catene contengono **almeno uno** dei due motivi?
Esempio: SLEQSE**RRRRL**LELQKS...

Soluzione

15 PC7, 20 PACE, 32 almeno uno dei due.