Programma python

Scrivere un programma che:

- prenda in ingresso un file che contiene un elenco di elementi regolatori (RBP e miRNA) divisi in clusters
- stampi i cluster ordinati per numero di elementi regolatori
- stampi per ogni cluster il numero di elementi regolatori, il numero di RBP (ed il numero di famiglie di RBP) ed il numero di mRNA (ed il numero di famiglie di mRNA)

Si può assumere che gli elementi regolatori di una stessa famiglia differiscono solo per l'ultimo carattere (e.g. IGF2BP1 e IGF2BP2 per le RBP, hsa-miR-130a e hsa-miR-130b per l'mRNA)

File di ingresso (`clusters`)

```
clustid regid   regname
1       352     IGF2BP1
1       353     IGF2BP2
1       354     IGF2BP3
2       8       AUF1
3       24      ELAVL1
4       383     PUM1
5       352     IGF2BP1
5       353     IGF2BP2
5       354     IGF2BP3
5       384     PUM2
6       84      hsa-miR-130a
6       85      hsa-miR-130b
6       104     hsa-miR-148a
6       203     hsa-miR-301a
6       204     hsa-miR-301b
6       352     IGF2BP1
6       353     IGF2BP2
6       354     IGF2BP3
7       376     PABP
8       114     hsa-miR-15a
...
```

Esempio esecuzione

```
>>> python reg_stats.py
Inserire nome file: clusters
clustid numreg  numrpb  nummrna
11      11      3(1)    8(1)
6       8       3(1)    5(3)
10      8       3(1)    5(5)
8       7       3(1)    4(3)
12      5       3(1)    2(1)
5       4       4(2)    0(0)
14      4       4(2)    0(0)
13      4       4(2)    0(0)
1       3       3(1)    0(0)
9       1       1(1)    0(0)
7       1       1(1)    0(0)
4       1       1(1)    0(0)
3       1       1(1)    0(0)
2       1       1(1)    0(0)
15      1       1(1)    0(0)
```

Programma python: suggerimento

Si possono implementare 5 funzioni separate:

1. una che legga il file ed estragga un dizionario id di gruppo (clustid), lista di elementi regolatori nel gruppo
2. una che data una lista di elementi regolatori, restituisca due liste, una di RBP ed una di mRNA (notare che il nome di un mRNA comincia sempre per `hsa`)
3. una che data una lista di RBP o di mRNA, restituisca il numero di famiglie in essa contenute
4. una che ordini i cluster per numerosità e ne stampi le statistiche usando le funzioni sopra
5. una (o un main) che realizzi il programma richiesto usando le funzioni di cui sopra

Esercizi da linea di comando

- Date le sequenze fasta contenute nella directory `fasta`, ciascuna descrivente una singola catena, stampare il nome del file e uno spazio seguito dai primi 3 amminoacidi delle catene che cominciano con un'alanina (A) o una fenilalanina (F), ordinate per sottosequenza.
- Risposta:

```
2h8ba.f AAA
1rmka.f ACS
...
1zr3d.f FTV
1msob.f FVN
```

Suggerimento

L'opzione `-k indice-colonna` permette di indicare a `sort` su quale colonna ordinare. Es. `sort -n -k 2` ordina numericamente basandosi sulla seconda colonna.

Esercizi da linea di comando

- Il file `clusters` contiene una lista di associazioni tra elementi regolatori (univocamente identificati da `regid` o `regname`) e cluster (identificati da `clustid`).
- Stampare la lista degli elementi regolatori preceduta dal numero di cluster in cui sono presenti, in ordine numerico crescente.
- Risposta:

```
1 AUF1
1 CUGBP1
1 ELAVL1
...

9 IGF2BP3
```