Esame 5/07/2016

Andrea Passerini
passerini@disi.unitn.it

Informatica

Programma python

Scrivere un programma python che:

- prenda in ingresso
  - un file con un elenco di famiglie PFAM e relative statistiche
  - un pattern che caratterizzi delle famiglie di cui si vogliono estrarre statistiche aggregate (e.g. `finger` per le zinc finger)
- restituisca le statistiche di ciascuna delle famiglie che soddisfano il pattern, e le statistiche medie su tutte queste famiglie

Esempio file

```
D          Accession        Type     Seq Num (seed)   Seq Num (full)   Avg length      Avg %id Avg co
z-alpha    PF02295 Domain   5        197      63.70   30       12.36   Adenosine deaminase z-alpha do
Z1         PF10593 Domain   66       197      229.90  27       27.47   Z1 domain
ZapA       PF05164 Family   445      1489     97.70   19       83.08   Cell division protein ZapA
ZapB       PF06005 Family   39       202      72.20   38       79.48   Cell division protein ZapB
ZapC       PF07126 Family   38       116      169.30  40       94.08   Cell-division protein ZapC
Zds_C      PF08632 Domain   11       275      51.50   65       5.59    Activator of mitotic machinery
Zea_mays_MuDR    PF05928 Family   4        1        115.00  100      34.53   Zea mays MURB-like pro
Zein       PF01559 Family   7        120      136.30  39       93.60   Zein seed storage protein
```

Esempio esecuzione

```
> python pfam_statistics.py
Inserire nome file: pfam_database
Inserire pattern da cercare: ribbon
D          Seq Num (seed)  Seq Num (full)  Avg length      Avg %id Avg coverage    Description
zf-RING_7        201     667     33.80   37      13.67   C4-type zinc ribbon domain
Zn_Tnp_IS1595    38      460     47.70   29      16.28   Transposase zinc-ribbon domain
zf-RING_9        122     723     185.30  34      26.82   Putative zinc-RING and/or ribbon
zf-NADH-PPase    55      1309    31.60   32      9.70    NADH pyrophosphatase zinc ribbon doma
zf-ribbon_3      28      306     25.40   37      9.74    zinc-ribbon domain
zinc_ribbon_11   2       8       125.90  33      93.67   Probable zinc-ribbon
zinc_ribbon_10   27      549     53.00   44      13.92   Predicted integral membrane zinc-ribb
zinc_ribbon_13   40      135     62.40   41      86.41   Nucleic-acid-binding protein containi
zinc_ribbon_12   48      356     44.50   42      5.78    Probable zinc-ribbon domain
zinc_ribbon_15   53      270     78.10   24      67.99   zinc-ribbon family
zf-LITAF-like    364     1131    68.20   27      35.64   LITAF-like zinc ribbon domain
zinc_ribbon_16   3       411     111.50  31      11.38   Zinc-ribbon like family
Zn_ribbon_17     9       414     55.60   36      5.06    Zinc-ribbon, C4HC2 type
Zn_ribbon_recom  325     2615    61.40   22      12.07   Recombinase zinc beta ribbon domain
zinc-ribbons_6   28      71      65.60   43      83.55   zinc-ribbons
zf-trcl 96       426     47.60   52      39.62   Probable zinc-ribbon domain
Zn-ribbon_8      580     1174    40.60   37      43.40   Zinc ribbon domain
zf-C2H2_8        3       61      93.20   62      22.09   C2H2-type zinc ribbon
Zn_ribbon_2      127     281     78.70   35      55.12   Putative zinc ribbon domain
zinc-ribbon_6    26      233     90.50   34      25.60   zinc-ribbon domain
zn-ribbon_14     77      490     31.80   58      8.59    Zinc-ribbon
zinc_ribbon_9    70      797     33.90   39      9.96    zinc-ribbon
zinc_ribbon_2    230     1543    22.90   40      7.49    zinc-ribbon domain
zinc_ribbon_5    3       150     36.80   50      9.01    zinc-ribbon domain
zinc_ribbon_4    57      324     36.00   34      9.43    zinc-ribbon domain
zinc_ribbon_6    132     874     58.10   45      11.23   Zinc-ribbon
AVG      105.54  606.85  62.31   38.38   28.20   average
```

Programma python: suggerimento

Si possono implementare 5 funzioni separate:

1. Una che legga il file e restituisca un dizionario da nome PFAM a statistiche+descrizione, e l'intestazione corrispondente
2. Una che dato il dizionario di famiglie ed un pattern, restituisca un dizionario con solo le famiglie le cui descrizioni contengano il pattern
3. Una che dato un dizionario di famiglie, restituisca una lista di valori medi per le statistiche associate alle famiglie
4. una che dato dizionario, statistiche medie e intestazione stampi l'informazione richiesta (vedi esempio)
5. una che realizzi il programma richiesto usando le funzioni di cui sopra

Shell: esercizio #1

Dato il file `pfam_database`, stampare a schermo tutti i domini che menzionano l'acetilene (identificati dalla stringa "`C2H2`") con copertura media maggiore o uguale al 20%.

Vanno stampati il nome Pfam del dominio (prima colonna) e la copertura media (penultima colonna), ordinati per copertura crescente.

Risultato atteso

```
zf-C2H2_8    22.09
zf-C2H2_2    24.20
zf-C2H2_aberr    26.25
```

Shell: esercizio #2

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

1. Una fenilalanina (F); un residuo qualunque; una fenilalanina o una tirosina (Y); una prolina (P). Il motivo può trovarsi in posizione arbitraria.
2. Una prolina (P); una treonina (T) o una serina (S); una alanina (A); un'altra prolina. Il motivo non deve trovarsi nè all'inizio nè alla fine della sequenza.
3. Il primo motivo seguito dal secondo, oppure il secondo seguito dal primo.

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

Risultato atteso

(1) 44 (2) 19 (3) 6

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