Esame 18/01/2017

Andrea Passerini
passerini@disi.unitn.it

Informatica

Programma python

Scrivere un programma python che:

- prenda in ingresso due file contenenti i risultati di due repliche di esperimenti di eCLIP (RBP binding sites su RNA) ed un valore di soglia
- stampi per ciascun RNA tutti e soli i siti di binding che appaiono in entrambe le repliche, considerando la sottosequenza minima comune tra i siti delle due repliche, solo se di lunghezza almeno pari alla soglia

Esempio file

Replica 1:

```
ENSG00000146963 372     426     SBDS_K562_rep01_1       1.2358810752525 +
ENSG00000164898 1274    1328    SBDS_K562_rep01_2       1.2358810752525 +
ENSG00000164638 26387   26392   SBDS_K562_rep01_3       4.5838043786728 +
ENSG00000164638 26392   26412   SBDS_K562_rep01_4       4.5838043786728 +
ENSG00000201041 44      81      SBDS_K562_rep01_5       1.67691378306429        +
ENSG00000196367 3818    3871    SBDS_K562_rep01_6       0.968468228908129       +
ENSG00000198839 82326   82368   SBDS_K562_rep01_7       1.7134396590894 +
ENSG00000127399 281     322     SBDS_K562_rep01_8       2.26187628378544        +
ENSG00000188707 75      133     SBDS_K562_rep01_9       2.63627179856694        +
ENSG00000127399 7225    7283    SBDS_K562_rep01_10      2.63627179856694        +
ENSG00000106682 21923   21967   SBDS_K562_rep01_11      2.63627179856694        +
...
```

Replica 2:

```
ENSG00000164638 26392   26411   SBDS_K562_rep02_1       5.2122922366814 +
ENSG00000164638 26387   26392   SBDS_K562_rep02_2       5.2122922366814 +
ENSG00000146701 16301   16362   SBDS_K562_rep02_3       3.02786766554397        +
ENSG00000146701 16812   16904   SBDS_K562_rep02_4       3.2122922366814 +
ENSG00000186480 137     181     SBDS_K562_rep02_5       2.76483325971018        +
ENSG00000146963 368     381     SBDS_K562_rep02_6       0.616862736909177       +
ENSG00000164898 1270    1283    SBDS_K562_rep02_7       0.616862736909177       +
ENSG00000196367 3819    3869    SBDS_K562_rep02_8       1.00402092358961        +
ENSG00000196262 3       39      SBDS_K562_rep02_9       3.4753266425152 +
ENSG00000186480 11584   11605   SBDS_K562_rep02_10      4.2122922366814 +
...
```

Esempio siti in comune

Replica 1:

```
...
ENSG00000189043 110     142     SBDS_K562_rep01_411     0.413879377230492       +
ENSG00000189043 164     210     SBDS_K562_rep01_417     0.0919512823431295      +
...
```

Replica 2:

```
...
ENSG00000189043 125     235     SBDS_K562_rep02_179     2.12482939543106        +
ENSG00000189043 6571    6606    SBDS_K562_rep02_217     2.62732973596025        +
ENSG00000189043 1375    1409    SBDS_K562_rep02_327     1.30540164107288        +
...
```

Siti in comune:

```
ENSG00000189043 125     142
ENSG00000189043 164     210
```

Esempio esecuzione

```
> python binding_sites.py
Nome file prima replica: SBDS-human_K562_ENCSR059CWF_rep1.bed
Nome file seconda replica: SBDS-human_K562_ENCSR059CWF_rep2.bed
Soglia su lunghezza sito: 10

ENSG00000164548 76      110

ENSG00000252542 29      54

ENSG00000200959 40      68

ENSG00000200408 52      69

ENSG00000252623 30      66

ENSG00000110107 15497   15541
...
```

Programma python: suggerimento

Si possono implementare 5 funzioni separate:

1. Una che legga un file e restituisca un dizionario id di RNA $\rightarrow$ elenco di suoi siti di binding
2. Una che dati due elenchi di siti di binding e una soglia, restituisca l'elenco di sotto-siti in comune di lunghezza pari almeno alla soglia
3. Una che dati due dizionari di siti corrispondenti a due diverse repliche, restituisca un dizionario di sotto-siti (usando la funzione precedente)
4. Una che stampi un dizionario di siti
5. una che realizzi il programma richiesto usando le funzioni di cui sopra, eventualmente sostituita con il main del programma

Shell: esercizio #1

Dato il file `sequences.fasta`, stampare a schermo gli identificativi delle sequenze che non cominciano per metionina (M), ordinati lessicograficamente. Preferibilmente usando una sola pipeline.
Sono accettabili eventuali linee di output vuote.

Risultato atteso

P02275 P05621 P06180 ... Q9I920 Q9LLA7
in totale l'output consiste di 31 righe non vuote

Shell: esercizio #2

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

- Una metionina (M) all'inizio della sequenza; seguita da due, tre o quattro aminoacidi qualunque; seguiti da una alanina (A), una glutammina (Q), o una metionina; seguita da due, tre o quattro aminoacidi qualunque; seguiti da una metionina.
- Una lisina (K) o una valina (V) alla fine della sequenza; preceduta da esattamente due aminoacidi qualunque; preceduti da una lisina o una valina; preceduta (a distanza arbitraria) da un acido aspartico (D), una serina (S), ed un acido aspartico.

Quante sequenze contengono almeno uno dei due motivi? E quante entrambi?

Risultato atteso

24, 6, 30, 0.

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