Esame 21/03/2013

Andrea Passerini
passerini@disi.unitn.it

Informatica

Programma python

Scrivere una funzione `averageMatchConservation(data)` che:

- prenda in ingresso un nome di file `data` che contiene un allineamento multiplo di domini proteici
- stampi i valori di conservazione di ciascuna posizione di match

File degli allineamenti (`alignments`)

```
# STOCKHOLM 1.0
#=GF ID   RRM_1
#=GF AC   PF00076.17
#=GF DE   RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
#=GF PI   rrm;
#=GF AU   Eddy SR, Birney E
....
A0CEN5_PARTE/256-312  [...].kemr------.[...]
A0CA51_PARTE/258-315  [...]..ekemr------.......-......-..[...]
A0C441_PARTE/141-196  [...]....T....Y..........G.........P.[...]
A0BF98_PARTE/253-307  [...].....T....Y..........G.........P...[...]
...
```

Esempio esecuzione

```
>>> import utility
>>> utility.averageMatchConservation('alignments')
(0, 0.816345761221206)
(1, 0.880013820397650 5)
(2, 0.913873794641454 9)
(3, 0.921851933285171 3)
(4, 0.926374972516254 7)
(5, 0.932091591544429 4)
(6, 0.934510161133272 6)
(7, 0.938970380375035 3)
(8, 0.944906869365832 2)
(9, 0.951157458303232 1)
(10, 0.955429217577033)
...
```

Programma python: suggerimento

Si possono implementare 5 funzioni separate:

1. una che legga il file e restituisca un dizionario con nome di dominio proteico come chiave, e suo allineamento come valore
2. una che dato un allineamento, restituisca una lista con 1 per le posizioni di match, e 0 per quelle di cancellazione (ignorando gli inserimenti)
3. una che scorra il dizionario di allineamenti, per ciascuno calcoli la lista di match e cancellazioni (funzione precedente), e calcoli la media di queste liste
4. una che stampi la media di match-cancellazioni
5. una che realizzi il programma richiesto usando le funzioni di cui sopra

Esercizio Shell 1

Calcolare il numero di domini proteici nel file `alignments` che cominciano con un inserimento

Soluzione

5620

Esercizio Shell 2

Contare tra i file fasta quante catene:

- cominciano con una alanina (A), seguita da un residuo qualunque e poi da una cisteina (C) o una istidina (H)
- e non finiscono con un acido (E o D)

Esempio:
AGCIKNGGRCNASAGPPYCCSSYCFQIAGQSYGVCKN

Soluzione

6