Esame 12/01/2012

Andrea Passerini
passerini@disi.unitn.it

Informatica

Programma python

Scrivere una funzione `composition4localization(fastafile)` che:

- prenda in ingresso un file `fastafile` con nome e localizzazione delle sequenze contenute
- legga l'insieme di sequenze dal file
- divida le sequenze per localizzazione
- calcoli la composizione aminoacidica delle sequenze in ciascuna localizzazione
- stampi per ogni localizzazione tale composizione (ordinata per frequenza)

Esempio dati

```
$ cat seq.fasta
>7B2_HUMAN:Secretory
MVSRMVSTMLSGLLFWLASGWTPAFAYSPRTPDRVSEADIQR
...
>68MP_BOVIN:Mitochondrion
MLQSLIKKVWIPMKPYYTQAYQEIWVGTG
...
```

Esempio esecuzione

```
>>> import utility
>>> utility.composition4localization('seq.fasta')
Mitochondrion
7147:L 5918:A 5016:V 4980:G 4766:E 4603:R ...
Cytoplasm
28198:L 24036:S 23509:E 21491:A 18761:K ...
Nucleus
69815:S 65844:L 56596:E 55254:A 52798:P ...
Secretory
24869:L 23092:S 21629:G 19221:A 18788:T ...
```

Programma python: suggerimento

Si possono implementare 5 funzioni separate:

1. una che legga da un file FASTA in ingresso le sequenze aminoacidiche e le restituisca (e.g. dizionario)
2. una che prenda in ingresso il dizionario di sequenze, e crei un dizionario di localizzazioni, ciascuna con una lista delle sequenze che hanno tale localizzazione
3. una che prenda in ingresso una lista di sequenze e ne restituisca la composizione aminoacidica
4. una che stampi il dizionario di frequenze (in ordine decrescente!! vedere funzioni per ordinare una lista)
5. una che realizzi il programma richiesto usando le funzioni di cui sopra

Esercizi da linea di comando

- Selezionare tra le sequenze in `seq.fasta` quelle che contengono almeno due volte consecutive un pattern fatto da:
  - due adenine (A)
  - da tre a dieci nucleotidi qualunque
  - due adenine (A) o tre timine (T)
- e.g. EERAANFENHAARLGAT
- e.g. TPTAAHKEATSTATTTATYA

Esercizi da linea di comando

- Calcolare il numero di sequenze che contengono una cisteina o un residuo che e' in contatto con una cisteina.
- Output: 73

Suggerimento

usare `grep` come primo comando della pipe