Esame 12/01/2012

Andrea Passerini
passerini@disi.unitn.it

Informatica

Programma python

Scrivere una funzione `findBindingPatternFreqs(sitefile, fastafile, k)` che:

- prenda in ingresso:
  - un file `sitefile` con siti di legame di una RBP
  - un file `fastafile` con le sequenze fasta in cui si trovano i siti
  - una lunghezza *k* per il sottosito da considerare (parte iniziale del sito)
- legga i due file
- estragga le sottosequenze corrispondenti a ciascun sito
- calcoli le frequenze di occorrenza dei primi *k* nucleotidi di ciascuna sottosequenza (sito)
- stampi le frequenze trovate

Esempio dati

```
$ cat binding_sites
target  length  start   end
uc002pja.1_5UTR 426     3       37
uc002sde.1_3UTR 4305    2572    2629
uc002sdm.2_3UTR 1638    1076    1100
uc001efm.2_5UTR 522     428     460
.....

$ cat seq.fasta
.....
>uc002pja.1_5UTR 426
AGACTGCGGGAGGACCCTGAGGGGCCAGGGGTAGCCATGA
GGCCTAGTCTGGGACGGAGCCTCGGGCTGGAGTAGCTTCG
GGGGCCTGGGCTGCCCCCTGGCTGTGGTCCGTGGGAAAGG
GCCCTGCTCAGGCGGGGCGGGCTGGGGAGGCCTCCGGGGA
.....
```

Esempio esecuzione

```
>>> import utility
>>> utility.findBindingPatternFreqs('binding_sites',\
... 'seq.fasta', 5)
GCGTT   2
AAATG   1
GCCCG   1
GCCCA   5
AAATC   2
GCCCC   7
AAATA   2
GCCCT   6
.....
```

Programma python: suggerimento

Si possono implementare 5 funzioni separate:

1. una che legga da un file FASTA in ingresso le sequenze nucleotidiche e le restituisca (e.g. dizionario)
2. una che prenda in ingresso il file di binding sites e ne estragga nome di sequenza, posizione iniziale e finale di ciascun sito
3. una che prenda in ingresso un dizionario di sequenze, un elenco di siti e un valore k e:
   - per ogni sito, recuperi la sottosequenza (da pos iniziale a finale) che lo caratterizza e ne estragga i primi k elementi
   - aggiorni un dizionario contenente frequenze di stringhe di k elementi
4. una che stampi il dizionario di frequenze
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:
  - tre timine (T)
  - tre nucleotidi qualunque
  - tre adenine (A)
- e.g. TTTTCCTAACTTTGAAAAATTTTGCAAATGTCTTA

Esercizi da linea di comando

- Contare nel file `seq.fasta` il numero di sequenze che correspondano rispettivamente a 3UTR e 5UTR
- Output:

```
2009 3UTR
 322 5UTR
```

Suggerimento

guardare di cosa e' composto il nome delle sequenze