

Esame 21/01/2014

Andrea Passerini
passerini@disi.unitn.it

Informatica

Programma python

Scrivere una funzione

`findOccurrences(pdbfile, conversione, pattern)` che:

- prenda in ingresso:
 - un nome di file `pdbfile` con una proteina in formato PDB
 - un nome di file `conversione` con una tabella di conversione per aminoacidi da tre a una lettera
 - una stringa di aminoacidi `pattern` (codice ad una lettera)
- e stampi il numero di volte in cui la stringa di aminoacidi `pattern` compare nella proteina letta da `pdbfile` (ignorando gli overlaps)

File 1A26

```

HEADER      TRANSFERASE                      16-JAN-98    1A26
TITLE       THE CATALYTIC FRAGMENT OF POLY(ADP-RIBOSE) POLYMERASE
...
SEQRES      1      361  ALA LEU THR VAL SER ALA GLY THR LYS SER LYS LEU ALA
SEQRES      2      361  LYS PRO ILE GLN ASP LEU ILE LYS MET ILE PHE ASP VAL
SEQRES      3      361  GLU SER MET LYS LYS ALA MET VAL GLU PHE GLU ILE ASP
SEQRES      4      361  LEU GLN LYS MET PRO LEU GLY LYS LEU SER LYS ARG GLN
SEQRES      5      361  ILE GLN SER ALA TYR SER ILE LEU ASN GLU VAL GLN GLN
SEQRES      6      361  ALA VAL SER ASP GLY GLY SER GLU SER GLN ILE LEU ASP
SEQRES      7      361  LEU SER ASN ARG PHE TYR THR LEU ILE PRO HIS ASP PHE
SEQRES      8      361  GLY MET LYS LYS PRO PRO LEU LEU SER ASN LEU GLU TYR
SEQRES      9      361  ILE GLN ALA LYS VAL GLN MET LEU ASP ASN LEU LEU ASP
...
SEQRES     28      361  LEU LYS PHE ASN TYR LYS THR SER LEU TRP
HET        CNA      200      27
HETNAM      CNA CARBA-NICOTINAMIDE-ADENINE-DINUCLEOTIDE
FORMUL      2  CNA      C22 H29 N7 O13 P2
FORMUL      3  HOH      *90 (H2 O1)
...

```

File conversione

ala A
arg R
asn N
asp D
cys C
glu E
gln Q
gly G
his H
ile I
leu L
lys K
met M
phe F
pro P
ser S
thr T
trp W
tyr Y
val V

Esempio esecuzione

```
>>> import utility  
>>> utility.findOccurrences('1A26','conversione','LL')  
7
```

Programma python: suggerimento

Si possono implementare 4 funzioni separate:

- 1 una che legga il file `pdbfile` e restituisca la sequenza dei residui della proteina (in formato a tre lettere)
- 2 una che legga il file `conversione` restituisca un dizionario codice a tre lettere – > codice a una lettera
- 3 una che prenda in ingresso la sequenza in formato a tre lettere, il dizionario di conversione e restituisca una stringa in formato ad una lettera
- 4 una che realizzi il programma richiesto usando le funzioni di cui sopra (incluso il calcolo del numero di occorrenze)

Shell: Esercizio 1

Il file `pdbfile` contiene (anche) informazioni sugli atomi della proteina (righe che iniziano per `ATOM`). In fondo a queste righe compare la lettera che rappresenta l'elemento chimico (C,N,O o S per questa proteina).

Calcolare il numero di atomi di cui la proteina e' fatta, divisi per elemento chimico

Risultato

```
1767 C
 466 N
 525 O
  11 S
```

Shell: Esercizio 2

Contare quante delle sequenze nella directory `fasta` contengono il seguente motivo:

- Due glicine (G) separate da due o tre acidi (D o E) seguiti da un numero qualunque (anche nessuno) di residui non acidi.

Risultato

7