

Esame 12/01/2010

Andrea Passerini
passerini@disi.unitn.it

Informatica

Funzione python

- Scrivere una funzione `longest(s1, s2)` che:
 - prenda in ingresso due stringhe `s1` ed `s2`
 - restituisca la sottostringa più lunga contenuta in entrambe
- E.g. (`longest.py`):

```
from longest import *  
>>> longest("AAABBBFSSDSS", "AAABCABBBFSS")  
'ABBBFSS'
```

Funzione python: suggerimento

- 1 Iterare sulla prima e sulla seconda stringa con due diversi indici
- 2 Per ogni possibile punto iniziale nella prima e nella seconda:
 - provare a costruire una stringa comune a partire da tali punti iniziali
 - fermarsi quando le due stringhe di ingresso differiscono
 - oppure si è raggiunta la fine di una delle due
- 3 Se la stringa costruita è più lunga della massima stringa costruita in precedenza, memorizzarla
- 4 Alla fine della doppia iterazione, restituire la stringa memorizzata

Funzione python: versione semplificata

- Per chi non riesca a creare la funzione, provare a creare una funzione

`longest_starting(str_search, str_init)` che restituisce la più lunga sottostringa di `str_search` che sia anche sottostringa *iniziale* di `str_init`.

- E.g.:

```
>>> longest_starting("AABBBCCCC", "BBBCHHFDGD")
'BBBC'
>>> longest_starting("CHHHHDDA", "DDACHHHHH")
'DDA'
```

Esercizi da linea di comando

- Selezionare da un file FASTA (`seq.fasta`, allegato) tutte le righe che contengano:
 - una sequenza di quattro tra cisteine, istidine, aspartati o glutamati (e.g. CHDD)
 - seguita da una sequenza di residui che non siano cisteine, istidine, aspartati o glutamati lunga da uno a sei residui (e.g. SA, OPA)
 - seguita da un'altra sequenza di due tra cisteine, istidine, aspartati o glutamati
- e.g. CEEEGTIWSYHC

Esercizi da linea di comando

- Dato un file (`list` allegato) con una serie di righe contenenti:
 - nome proteina + “_” se proteina a catena singola (e.g. `1a6f_`)
 - nome proteina + lettera maiuscola indicante la catena se proteina multicatena (e.g. `1b71A`)
- prendere:
 - solo le proteine multicatena
 - estrarne i nomi rimuovendo l’indicazione della catena
 - stampare l’elenco delle proteine distinte eliminando la ridondanza se una proteina appare con più catene

Esercizi da linea di comando

- Output:

117e

11as

13pk

16gs

16vp

....

- Suggestione: `cut` ha un'opzione per selezionare un numero di caratteri invece che un numero di campi (e.g. `abc` sono 3 caratteri, `a b c` sono 3 campi separati da spazi)