Esame 12/07/2011

Andrea Passerini
passerini@disi.unitn.it

Informatica

Programma python

Date:

- Una directory (`Fasta`) contenenti sequenze proteiche in formato FASTA
- Una directory (`Profiles`) contenenti matrici di conservazione evolutiva delle sequenze. Una riga per ogni residuo della sequenza, contenente la frazione di volte in cui gli aminoacidi `VLIMFWYGAPSTCHRKQEND` si trovano nella posizione corrispondente nel profilo evolutivo

Programma python

Scrivere una funzione `prinConsSequence(fastafile, profilefile, residues, threshold)` che:

- prenda in ingresso i nomi del file fasta, del corrispondente file con i profili, dell'ordine degli aminoacidi nel profilo (ossia `VLIMFWYGAPSTCHRKQEND`) e della soglia di conservazione
- legga la sequenza proteica ed il suo profilo
- stampi la sequenza con lettera minuscola per i residui con conservazione sotto la soglia, maiuscola altrimenti
- E.g. (`utility.py`):

```
>>> import utility
>>> utility.printConsSequence('Fasta/1eptA',
... 'Profiles/1eptA', , 0.5)
IVGGytcaansiPyQVSLnsGsHfCGGSLInsqWVvsAAHCyk
```

Programma python: suggerimento

Si possono implementare 5 funzioni separate:

1. una che legga da un file FASTA in ingresso una sequenza proteica e la restituisca
2. una che legga da un file di profili in ingresso una matrice di conservazione e la restituisca
3. una funzione che prenda in ingresso un residuo, il suo profilo (corrispondente riga della matrice), l'ordine degli aminoacidi nella riga (ossia `VLIMFWYGAPSTCHRKQEND`) e la soglia di conservazione, restituisca vero se il residuo e' conservato, falso altrimenti.

Programma python: suggerimento (continua)

4. una funzione che prenda in ingresso la sequenza proteica, la matrice di profili, la (solita) stringa `VLIMFWYGAPSTCHRKQEND` e la soglia di conservazione e, usando la funzione 3 sopra, restituisca una stringa con la sequenza con residui non conservati minuscoli e conservati maiuscoli.
5. una funzione che utilizzi le quattro funzioni precedenti per produrre il risultato voluto

Esercizi da linea di comando

- Selezionare tra i file nella directory `Fasta` tutte le sequenze:
  - che comincino per un residuo che non sia una Metionina (M), seguito da tra due e cinque residui qualunque seguiti da una Cisteina (C)
  - oppure che finisca con una Cisteina (C) seguita da tra due e cinque residui qualunque
- e.g. VVIGQRCYRSPDCYSACKKLVGKATGKCTNGRCDC
- e.g. ...TVPVSHHECSFLKPCLCVSQRS

Esercizi da linea di comando

- Calcolare per un qualunque file nella directory `Profiles` le posizioni contenenti Cisteine (colonna 13 del profilo) con conservazione almeno 0.5
- Esempio di output per la proteina `1dnv_`:

```
218 0.800
232 0.566
384 0.540
386 0.730
```