Esame 15/09/2017

Andrea Passerini
andrea.passerini@unitn.it

Informatica

Programma python

Dati:

- Un file (`seqs.fasta`) contenente sequenze proteiche in formato FASTA
- Una directory (`Profiles`) contenenti matrici di conservazione evolutiva delle sequenze. Una riga di intestazione (`VLIMFWYGAPSTCHRKQEND`) con gli aminoacidi usati nelle righe successive. Una riga per ogni residuo della sequenza, contenente la frazione di volte in cui ciascuno degli aminoacidi `VLIMFWYGAPSTCHRKQEND` si trova nella posizione corrispondente nel profilo evolutivo
- Un valore di conservazione tra 0 e 1.

Programma python

Scrivere un programma `extractVariants` che:

- prenda in ingresso il file fasta, la directory di profili e il valore di conservazione
- per ogni sequenza, ne stampi il nome e l'elenco di residui, aggiungendo per ogni residuo le possibili alternative che risultano avere una conservazione maggiore del valore in ingresso

Esempio Dati di Ingresso

file fasta

```
...
>1bgk_
VCRDWFKETACRHAKSLGNCRTSQKYRANCAKTCELC
>1bjx_
AATTLPDGAAAESLVESSEVAVIGFFKDVESDSAKQFLQAAEAIDDIPFGITSNSD...
...
```

profilo

```
cat Profiles/1bjx_
```

Esempio esecuzione

```
> python extractVariants.py
Nome file fasta: seqs.fasta
Nome directory profili: Profiles
Soglia di conservazione: 0.4
16vpA SR[M|L]PSPPM[P|A][V|A][P|S]PAAL[F|Y]...
1a3c_ MNQKAVILDEQAIRRALTRIAHE[M|I]IERNKGMNN[C|L]...
1a4rA MQTIKCVVVGD[V|G][A|G]VGKTCLLIS[Y|F]TTNKFPSEYVPT...
1a5nA MELTPREKDKLL[L|I]FTA[A|G]L[V|L]AERRLARGLKLNYPE[S|A]...
1aa7A MSLLTEVETYVLSI[I|V]PSGPLKAEIAQRLEDVFAGKNTDLE[V|A]LM...
...
```

Programma python: suggerimento

Si possono implementare 4 funzioni separate:

1. una che legga il file fasta e restituisca una mappa nome $\rightarrow$ sequenza
2. una che data una sequenza, il nome del file di profili corrispondente e un valore di conservazione, legga dal file di profili l'elenco di aminoacidi, e poi scorra il file parallelamente alla sequenza in ingresso, tenendo per ogni riga il residuo della sequenza e gli altri aminoacidi con conservazione superiore alla soglia (se ce ne sono), e restituisca la stringa ottenuta concatenando quanto estratto
3. una che data la mappa di sequenze, la directory dei profili e il valore di conservazione, per ogni sequenza chiami la funzione precendente e stampi il risultato
4. una (o un main) che realizzi il programma richiesto usando le funzioni di cui sopra

Esercizi da linea di comando

- Calcolare quante sequenze del file `seqs.fasta` contengono questo pattern:
  - una metionina (M) ad inizio sequenza
  - un acido aspartico (D) o glutammico (E) entro cinque residui dalla fine della sequenza, ma non come ultimo residuo della sequenza

Risultato atteso

15

Esercizi da linea di comando

- Usando l'informazione nella directory `Profiles`, riportare i nomi delle cinque sequenze più lunghe

Risultato atteso

```
1i1iP
1n7uA
1fp4B
1dl2A
1ibrB
```

Modalita' di esecuzione e consegna

1. Avviare la macchina in modalita' `ESAME`
2. Autenticarsi con nome utente `sci-esame` e password fornita dal docente
3. Il testo del compito ed i file necessari si trovano in una cartella `Testo` sul Desktop
4. Realizzare il programma python come file `programma.py` e scrivere gli esercizi da linea di comando in un file di testo `linea_di_comando.txt`
5. Creare sul Desktop una cartella con *nome_cognome* e metterci i due file realizzati.
6. Eseguire il logout ma NON spegnere la macchina