Progetto

Laboratorio di Programmazione II

Corso di Laurea in Bioinformatica

Dipartimento di Informatica - Università di Verona

Progetto e prova finale

Istruzioni

Il codice realizzato durante il corso dovra' essere utilizzato nella prova finale al calcolatore. In quel contesto verranno forniti dei file fasta di prova e delle domande a cui si deve rispondere. Le domande saranno del tipo: quanti record contiene il file ? Quanti ORF nel frame 2 sono contenuti nel file, etc. Lo studente potra' utilizzare il codice realizzato durante il corso per rispondere a queste domande.

Progetto

Realizzare un modulo per analisi di FASTA file

Il modulo usa fastautil.py ed in particolare la funzione *read_fasta* che legge tutte le entry del file in un dizionario associando agli id le sequenze di DNA corrispondenti.

Funzionalita' richieste:

- P1 Contare il numero di record;
- P2 Analizzare la lunghezza delle stringhe;
- P3 Analizzare gli ORF (Oper Reading Frame);
- P4 Analizzare i *repeats* (stringhe ripetute di lunghezza data);

Scaricare e modificare il file progetto.py, realizzando tutte le funzioni come indicato nel modulo.
Il modulo ha un main di test che utilizza i file dna.simple.fasta e dna.long.fasta (scaricare i file per il corretto funzionamento).

P1: contare il numero di record

Spiegazione

Dato un file in formato FASTA, contare il numero di id nel file di input.

P2: Analizzare la lunghezza delle stringhe

Spiegazione

Dato un file in formato FASTA, creare un dizionario che associ gli id delle sequenze alla lunghezza delle sequenze stesse in modo da poter rispondere a domande quali: calcolare la lunghezza della sequenza piu' lunga e piu' corta, calcolare il numero di sequenze che hanno lunghezza massima ed il numero di sequenze che hanno lunghezza minima, calcolare la lista degli id associati alle sequenze massime e la lista degli associati alle sequenze minime.

P3: analizzare gli ORF

Spiegazione

Un reading frame e' una maniera di suddividere una sequenza di nucleotidi in triplette che non si sovrappongono. Data una sequenza ci sono 3 possibili reading frame 0 (che parte dal primo carattere) 1 (dal secondo) e 2 (dal terzo).Dato una sequenza ed un reading frame, un ORF e' una sottostring che inizia con ['ATG'] e temina con un codone di stop ['TGA','TAG','TAA'].

Dato un file in formato FASTA ed un frame di lettura (i.e., 0,1,2), creare un dizionario che associ gli id delle sequenze alla lista degli ORF contenuti in ciascuna sequenza, in modo da poter rispondere a domande quali: calcolare la lunghezza dell'ORF piu' lungo, calcolare l'id che contiene l'ORF piu' lungo.

P4: analizzare i *repeat*

Spiegazione

Un *repeat* e' una sottostringa di una sequenza di DNA che occorre almeno una volta nella sequenza.

Dato un file in formato FASTA ed un intero n, creare un dizionari che associ gli id delle sequenze con tutte i *repeat* di lunghezza n che occorrono in ciascuna sequenza, in modo da rispondere a domande quali: restituire la sottostringa che si ripete piu' di frequente e la frequenza con cui si ripete, calcolare l'insieme di sottostringhe che compaiono almeno k volte.