



Laboratorio di Elementi di Bioinformatica

Laurea Triennale in Informatica
(codice: E3101Q116)

AA 2018/2019

Esercizio (29 novembre 2018)

Docente del laboratorio: Raffaella Rizzi

[Esercizio]

Produrre uno script Python che prenda in input un file in formato GTF che annota un set di geni sulla stessa genomica di riferimento e il file FASTA della genomica di riferimento, e produca in standard output (in formato FASTA) le sequenze dei trascritti e delle CDS dei geni annotati nel file GTF in input.

Nell'header FASTA di ogni sequenza prodotta (trascritto o CDS) deve essere specificato:

- l'ID della genomica di riferimento
- l'ID del gene di riferimento
- l'ID del trascritto di riferimento
- il tipo di sequenza (cioé se è trascritto o CDS)
- la lunghezza della sequenza
- lo strand rispetto alla genomica di riferimento
- (solo nel caso di CDS) la presenza dello start e dello stop codon