Laboratorio di Elementi di Bioinformatica

Laurea Triennale in Informatica
(codice: E3101Q116)

AA 2016/2017

Formato GTF per annotare un gene

Docente del laboratorio: Raffaella Rizzi

GTF (Gene Transfer Format)

Il formato GTF ha lo scopo di annotare i trascritti (isoforme) di un gene su una sequenza genomica di riferimento, in termini di:

- ✓ composizione in esoni
- ✓ regioni non tradotte al 5’ (5’ UTR)
- ✓ regioni non tradotte al 3’ (3’ UTR)
- ✓ coding sequence (CDS)
- ✓ start e stop codon

GTF (Gene Transfer Format)

Il formato GTF ha lo scopo di annotare i trascritti (isoforme) di un gene su una sequenza genomica di riferimento, in termini di:

- ✓ composizione in e
- ✓ regioni non tra
- ✓ regioni non tradotte
- ✓ coding sequence (CDS)
- ✓ start e stop codon

… che deve contenere il *locus* del gene

GTF (Gene Transfer Format)

Il formato GTF ha lo scopo di annotare i trascritti (isoforme) di un gene su una sequenza genomica di riferimento, in termini di:

- ✓ composizione in e
- ✓ regioni non tra
- ✓ regioni non tradotte
- ✓ coding sequence (CDS)
- ✓ start e stop codon

… che deve contenere il *locus* del gene

Da un file GTF si possono quindi ricostruire tutti i trascritti (o isoforme) di un gene

GTF (Gene Transfer Format)

Il formato GTF ha lo scopo di annotare i trascritti (isoforme) di un gene su una sequenza genomica di riferimento, in termini di:

- ✓ composizione i[illegible]
- ✓ [illegible]
- ✓ [illegible]
- ✓ coding [illegible]
- ✓ start e stop codon

Attenzione! La sequenza presa come riferimento non deve essere necessariamente presa sulla catena di trascrizione del gene

GTF (Gene Transfer Format)

Il formato GTF:

- ✓ è un formato di puro testo che deriva dal formato GFF (General Feature Format)

GTF (Gene Transfer Format)

Il formato GTF:

- ✓ è un formato di puro testo che deriva dal formato GFF (General Feature Format)
- ✓ un file in formato GTF ha estensione `*.gtf` oppure `*.gff`

GTF (Gene Transfer Format)

Il formato GTF:

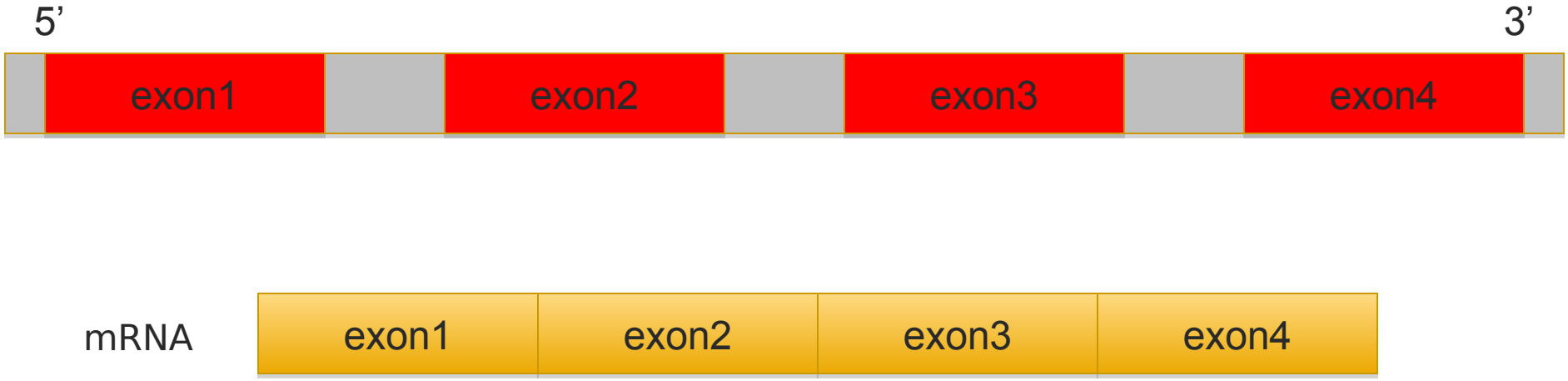
- ✓ è un formato di puro testo che deriva dal formato GFF (General Feature Format)
- ✓ un file in formato GTF ha estensione `*.gtf` oppure `*.gff`
- ✓ è composto da *record* di nove campi separati da tabulazione

GTF (Gene Transfer Format)

Ogni *record* descrive una *feature*, cioé una sottostringa della sequenza genomica di riferimento che rappresenta uno dei seguenti “oggetti”:

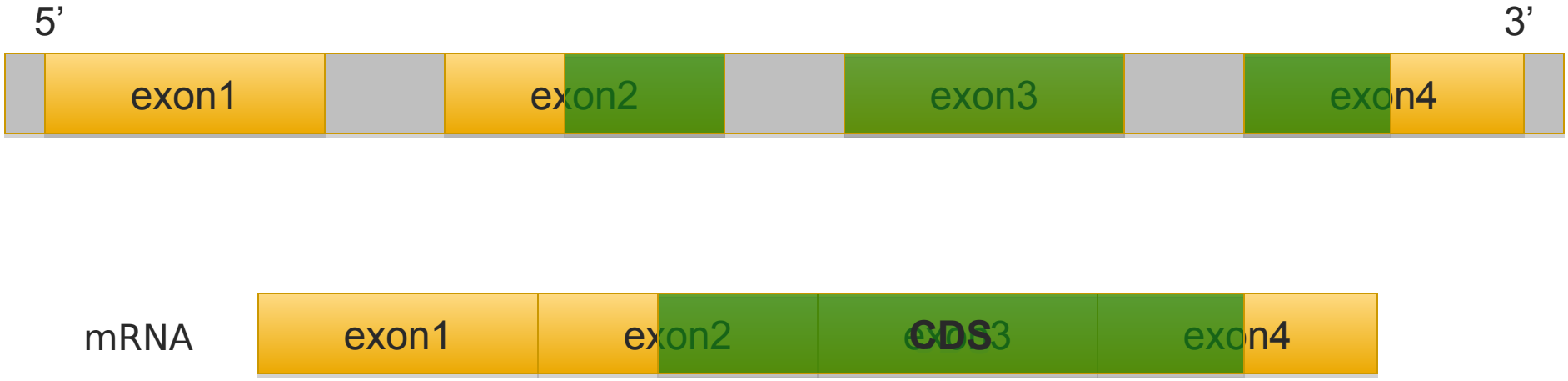
- ✓ un esone
- ✓ una CDS (in genere una sua parte)
- ✓ un 5’ UTR (in genere una sua parte)
- ✓ un 3’ UTR (in genere una sua parte)
- ✓ uno start codon
- ✓ uno stop codon

Feature relativa a un esone

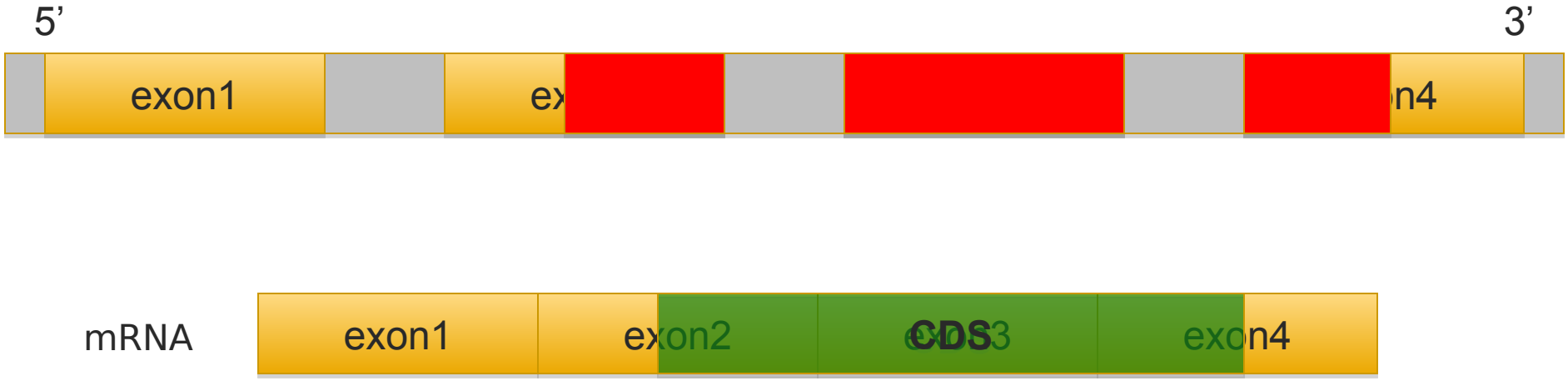


Ad ogni esone di un trascritto corrisponde una *feature* sulla sequenza di riferimento, e quindi un record nel formato GTF.

Features relative a una CDS

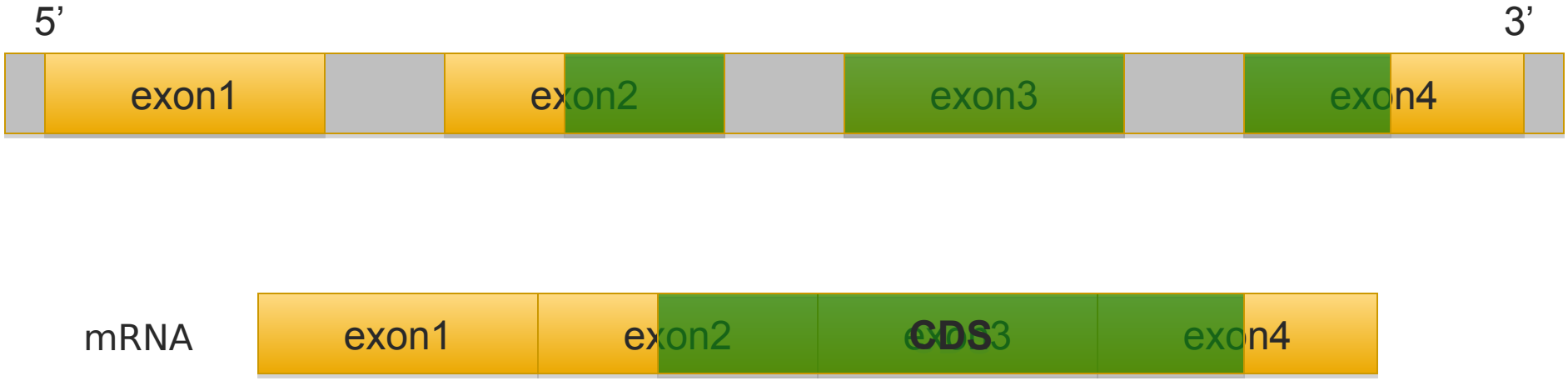


Features relative a una CDS

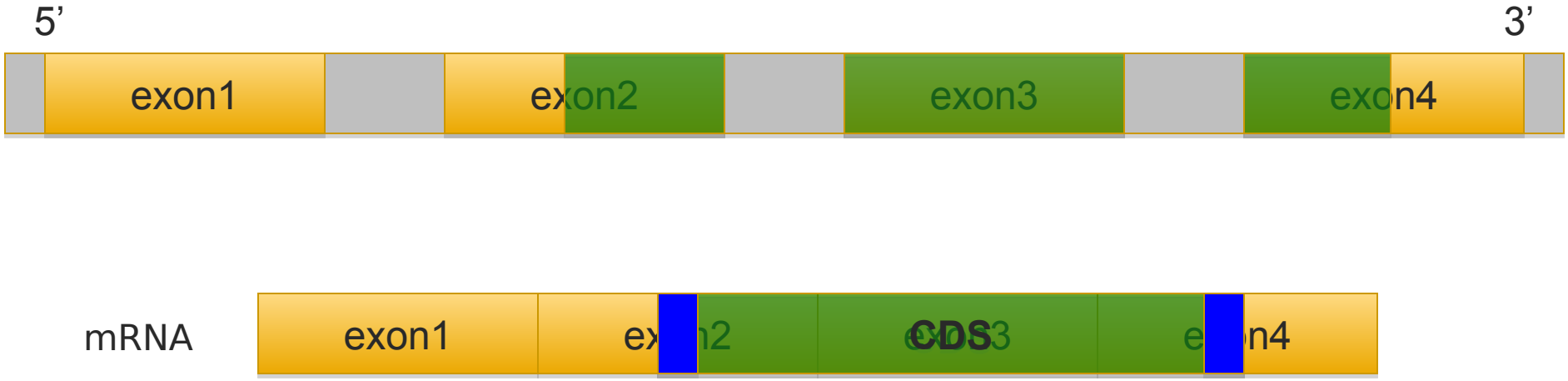


Alla CDS di un trascritto corrispondono una o più *features* sulla sequenza di riferimento, e quindi uno o più record nel formato GTF.

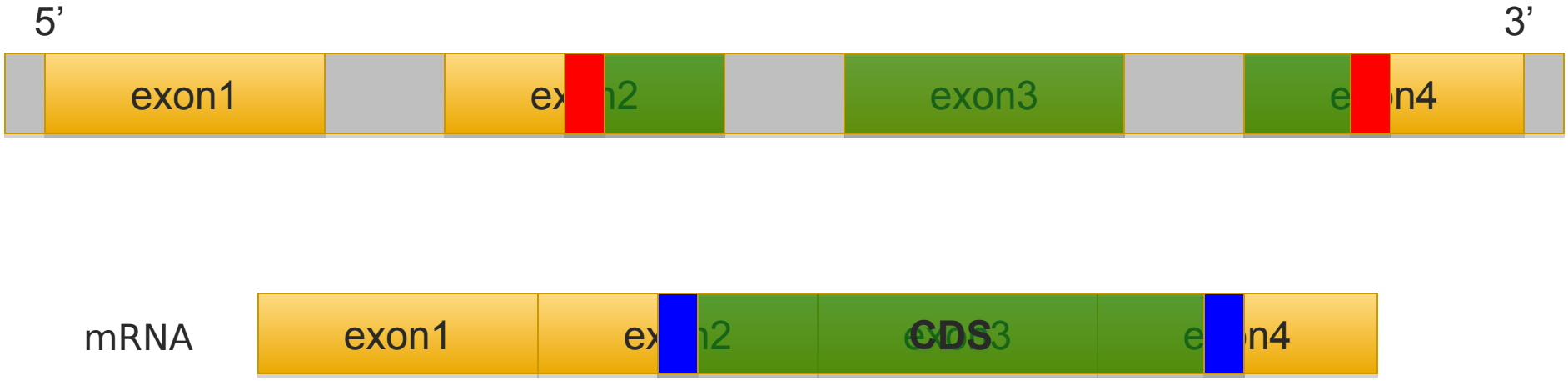
Features relative a start/stop



Features relative a start/stop

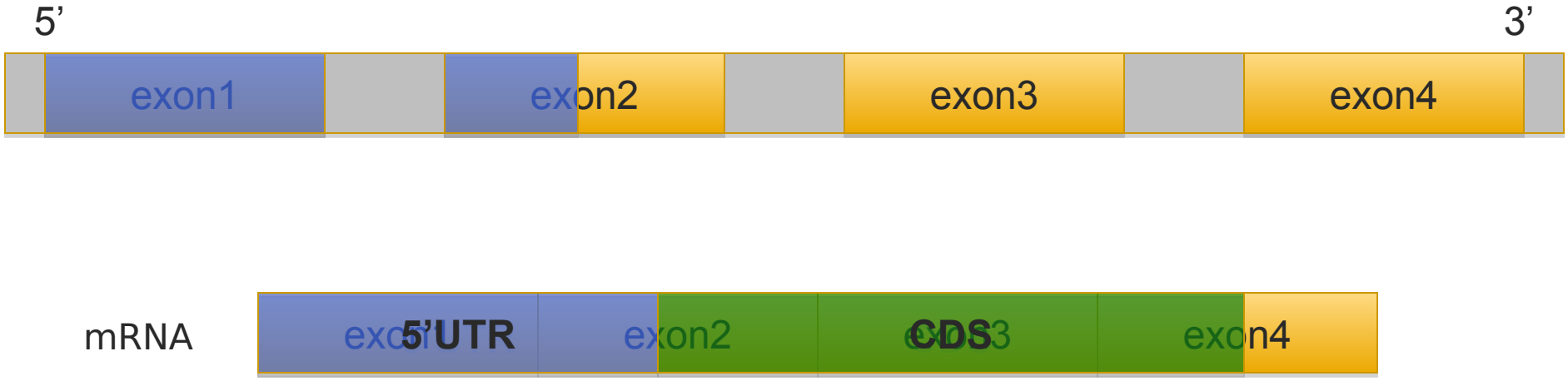


Features relative a start/stop

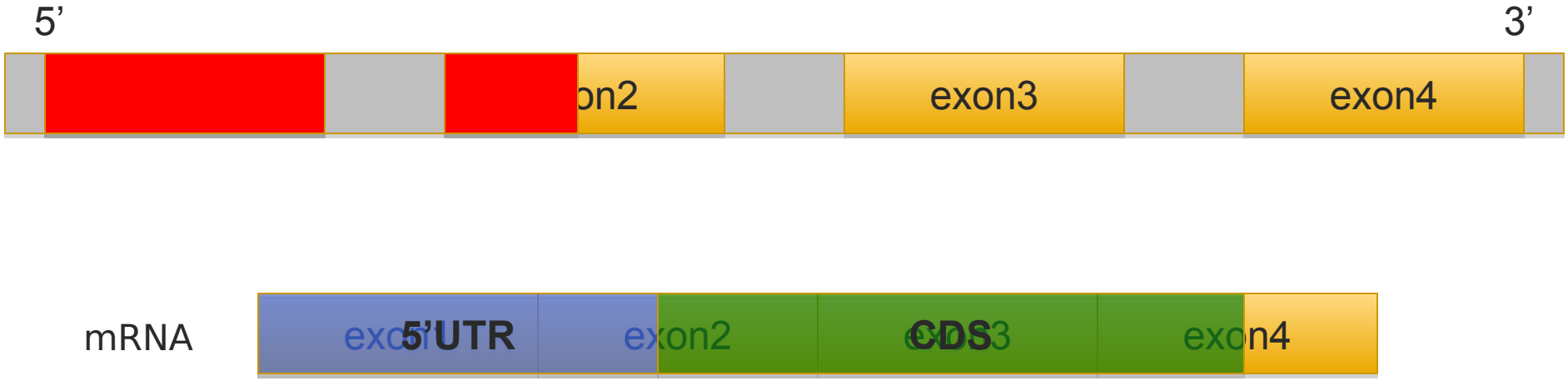


Ad uno start/stop codon di un trascritto corrispondono (in teoria) una o più *features* sulla sequenza di riferimento, e quindi uno o più record nel formato GTF.

Features relative a un 5’ UTR

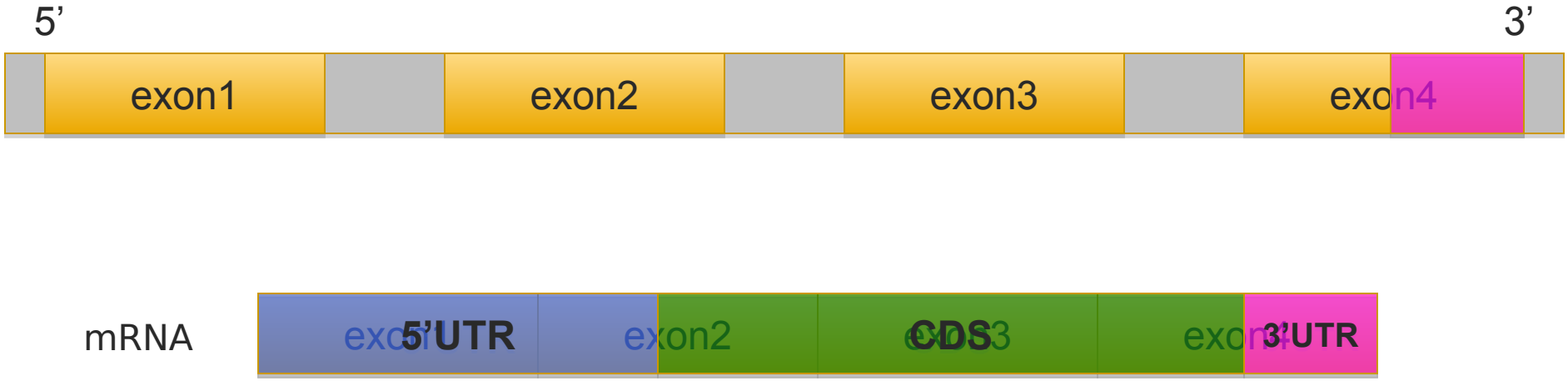


Features relative a un 5’ UTR

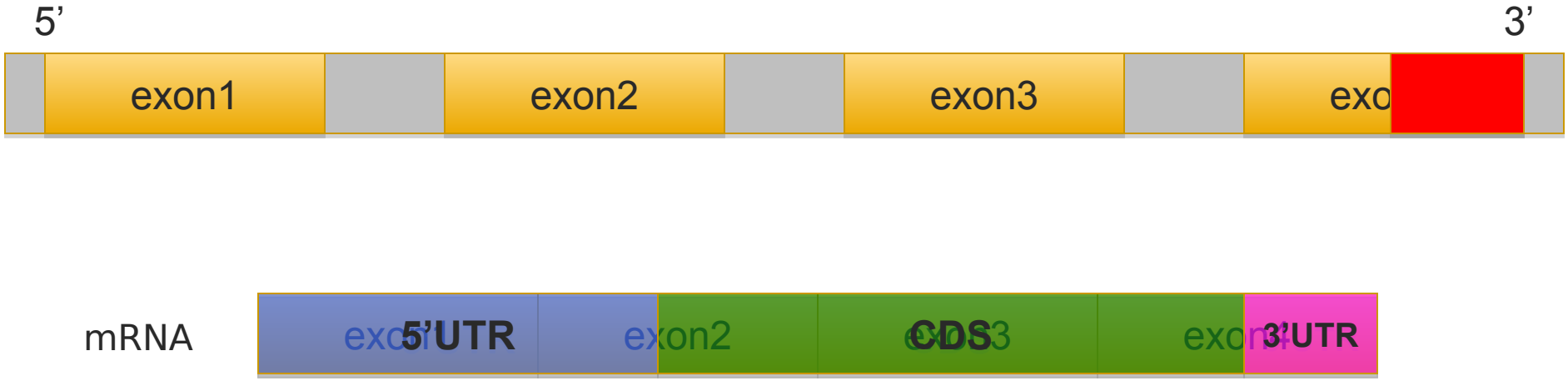


Al 5’ UTR di un trascritto corrispondono una o più *features* sulla sequenza di riferimento, e quindi uno o più record nel formato GTF.

Features relative a un 3’ UTR



Features relative a un 3’ UTR



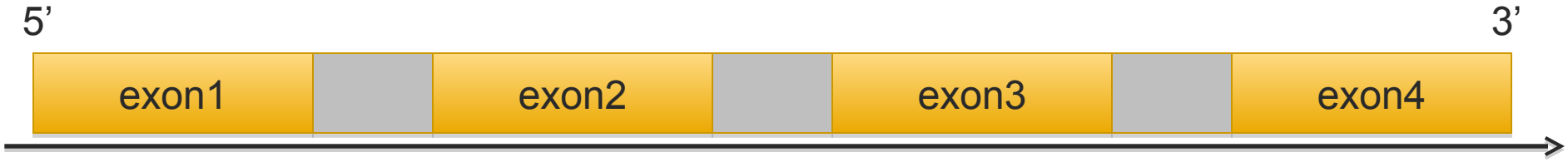
Al 3’ UTR di un trascritto corrispondono una o più *features* sulla sequenza di riferimento, e quindi uno o più record nel formato GTF.

I campi di un *record* GTF

Ogni *record* del formato GTF è composto dai seguenti nove campi:

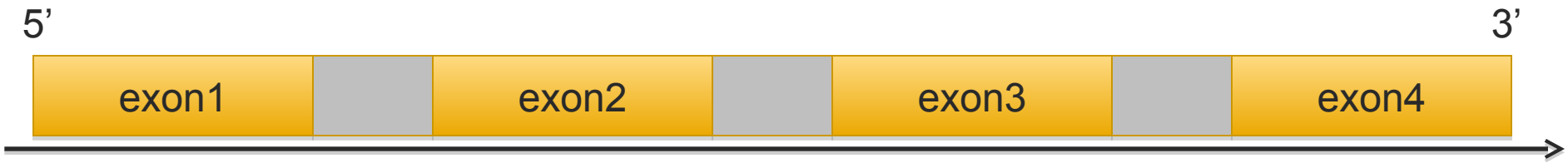
- ✓ identificatore della genomica di riferimento
- ✓ la sorgente che ha prodotto l'annotazione (ad esempio un *software*)
- ✓ il nome della *feature* (valori possibili "exon", "CDS", "5UTR", "3UTR", "start_codon", "stop_codon")
- ✓ la posizione di inizio della *feature* sulla genomica di riferimento
- ✓ la posizione di fine della *feature* sulla genomica di riferimento
- ✓ lo *score* della *feature* ('.', se alla *feature* non è associato uno *score*)
- ✓ lo *strand* (valori possibili +, -)
- ✓ il *frame,* solo per CDS, start e stop codon (valori possibili 0, 1, 2); se la *feature* non ammette un valore di frame si trova '.'
- ✓ il campo degli attributi nella forma:
  - ✓ `<attribute_name1> <value1>; <attribute_name2> <value2>; …`
  - ✓ Attributi obbligatori sono: "`gene_id`" e "`transcript_id`"

Lo strand

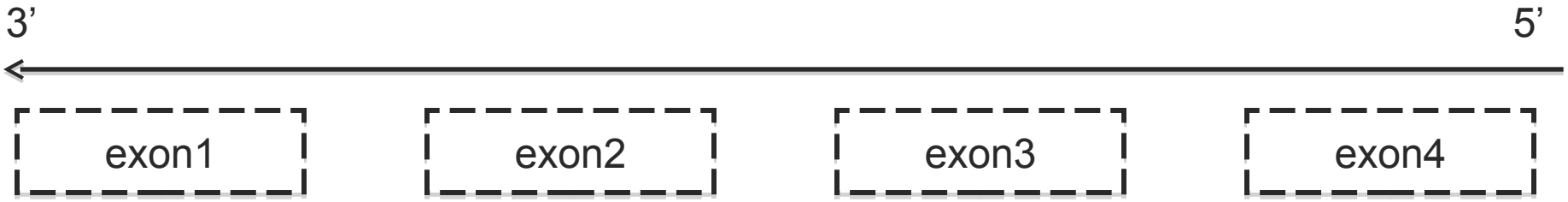


Annotare un gene su di una sequenza di riferimento presa sulla sua catena di trascrizione implica uno *strand* +

Lo strand



Annotare un gene su di una sequenza di riferimento presa sulla sua catena di trascrizione implica uno *strand* +



Annotare un gene su di una sequenza di riferimento presa sulla catena opposta alla sua catena di trascrizione implica uno *strand* -

Lo strand

Per ottenere la sequenza di una *feature* con *strand* -, si deve:

① estrarre la sottostringa della genomica di riferimento che corrisponde alla *feature*

②eseguire un'operazione di reverse&complement della sottostringa estratta

Esercizio

Scrivere un programma che prenda in input un file in formato GTF e il file FASTA della (unica) genomica di riferimento, e produca in standard output (in formato FASTA) le sequenze dei trascritti e delle CDS dei geni annotati nel file in input.

Nell'header FASTA di ogni sequenza prodotta deve essere specificato:

- l'ID della genomica di riferimento (primo campo del GTF)
- il nome 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
- (nel caso di CDS) la presenza dello start e dello stop codon