

Corso di Elementi di Bionformatica

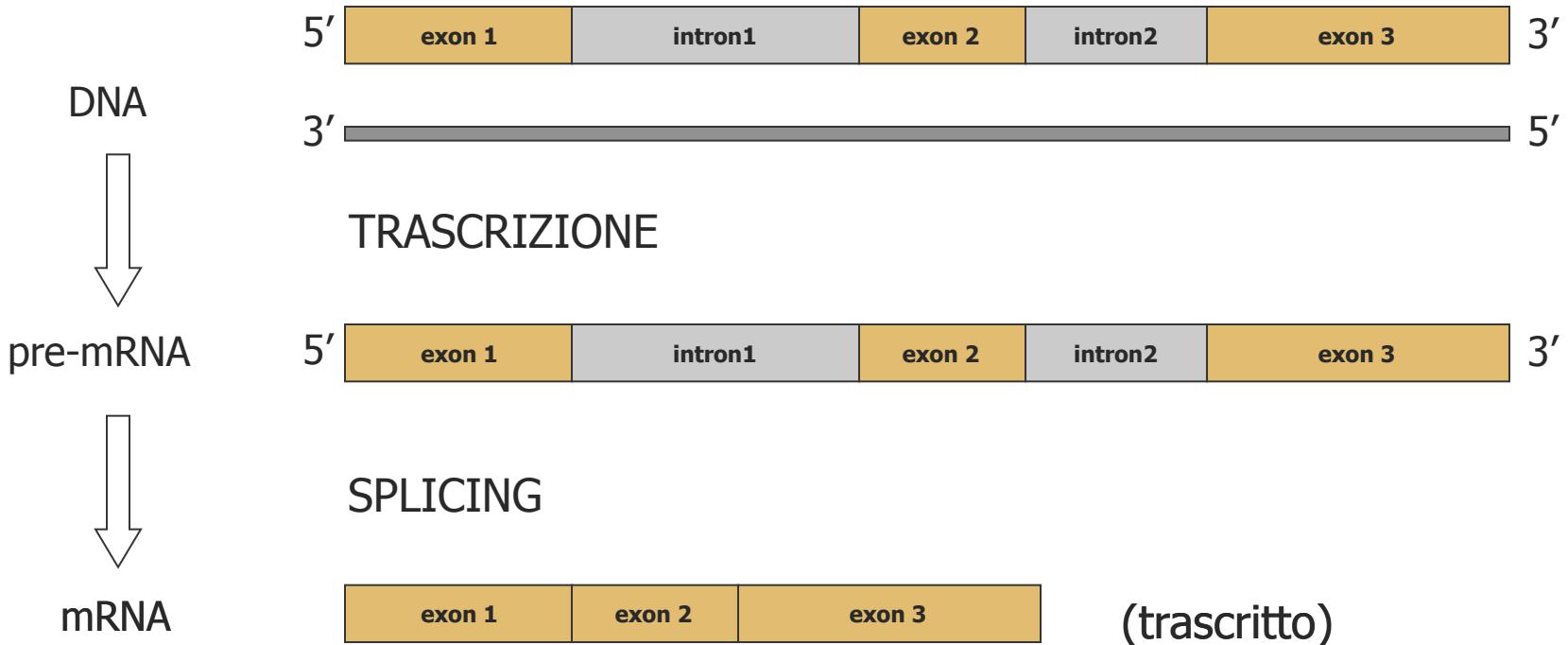
Laurea Triennale in Informatica

**Il formato GTF
per l'annotazione di un gene**

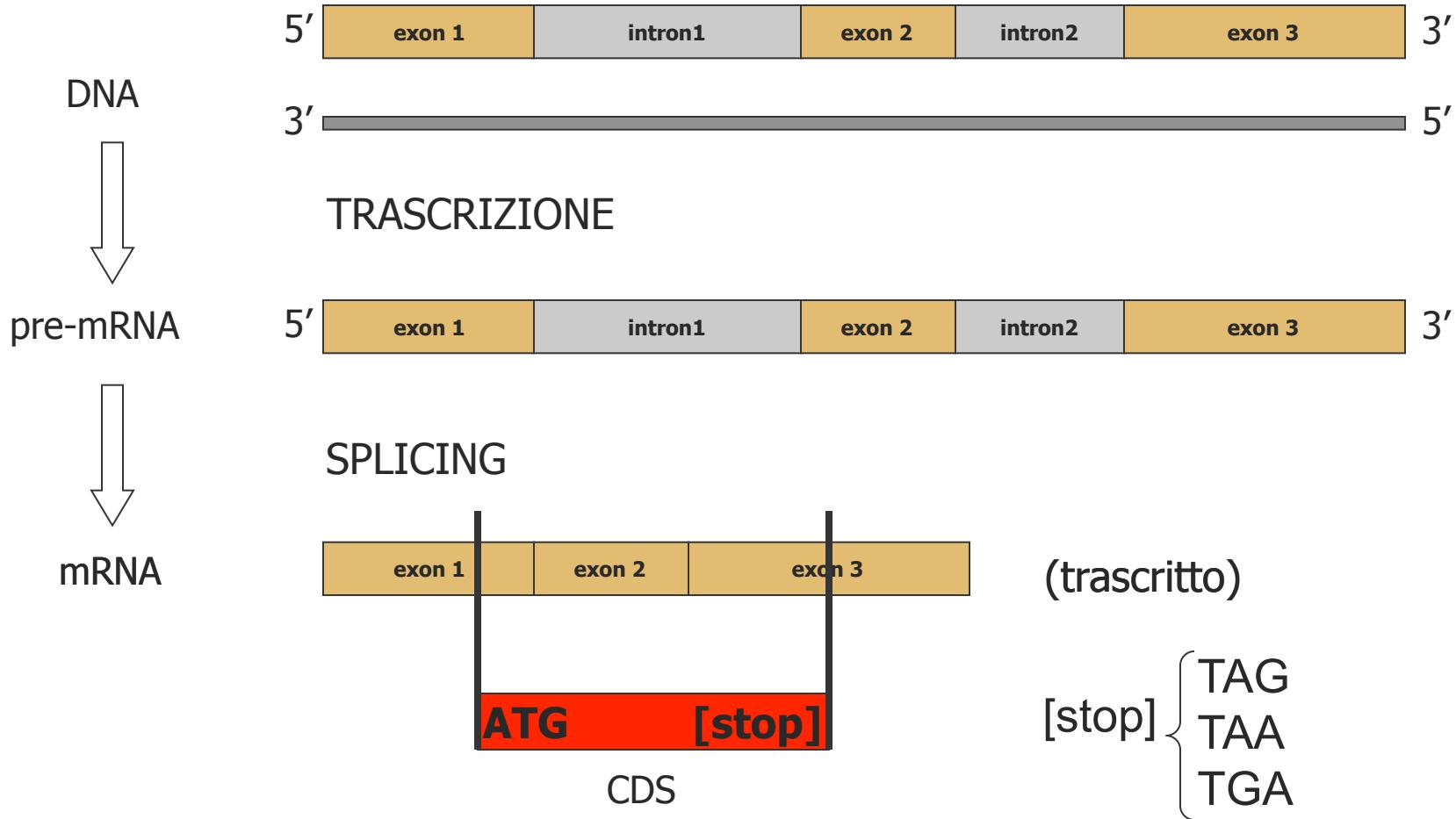
Anno Accademico 2015-2016

Docente del laboratorio: Raffaella Rizzi

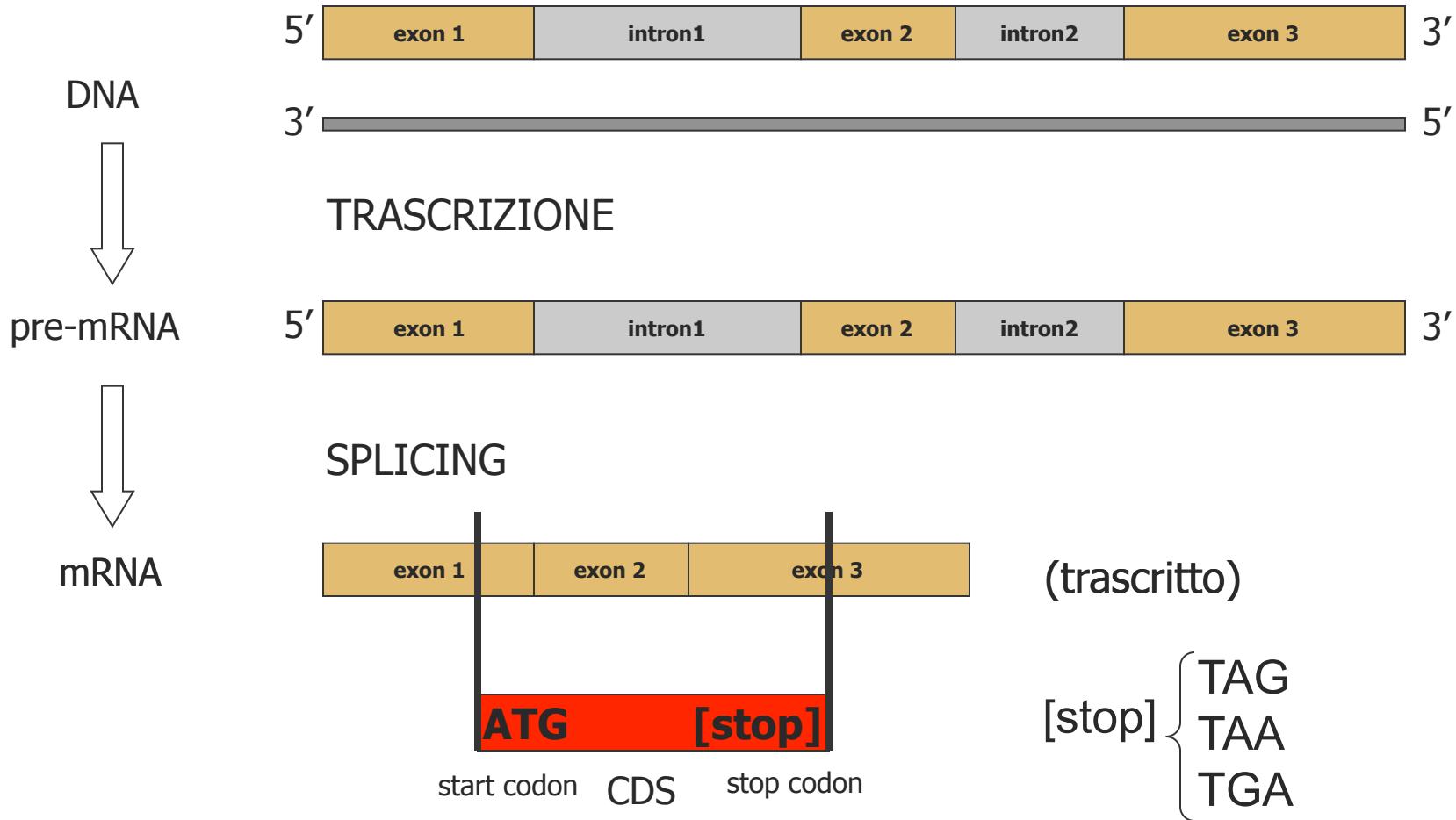
Espressione di un gene



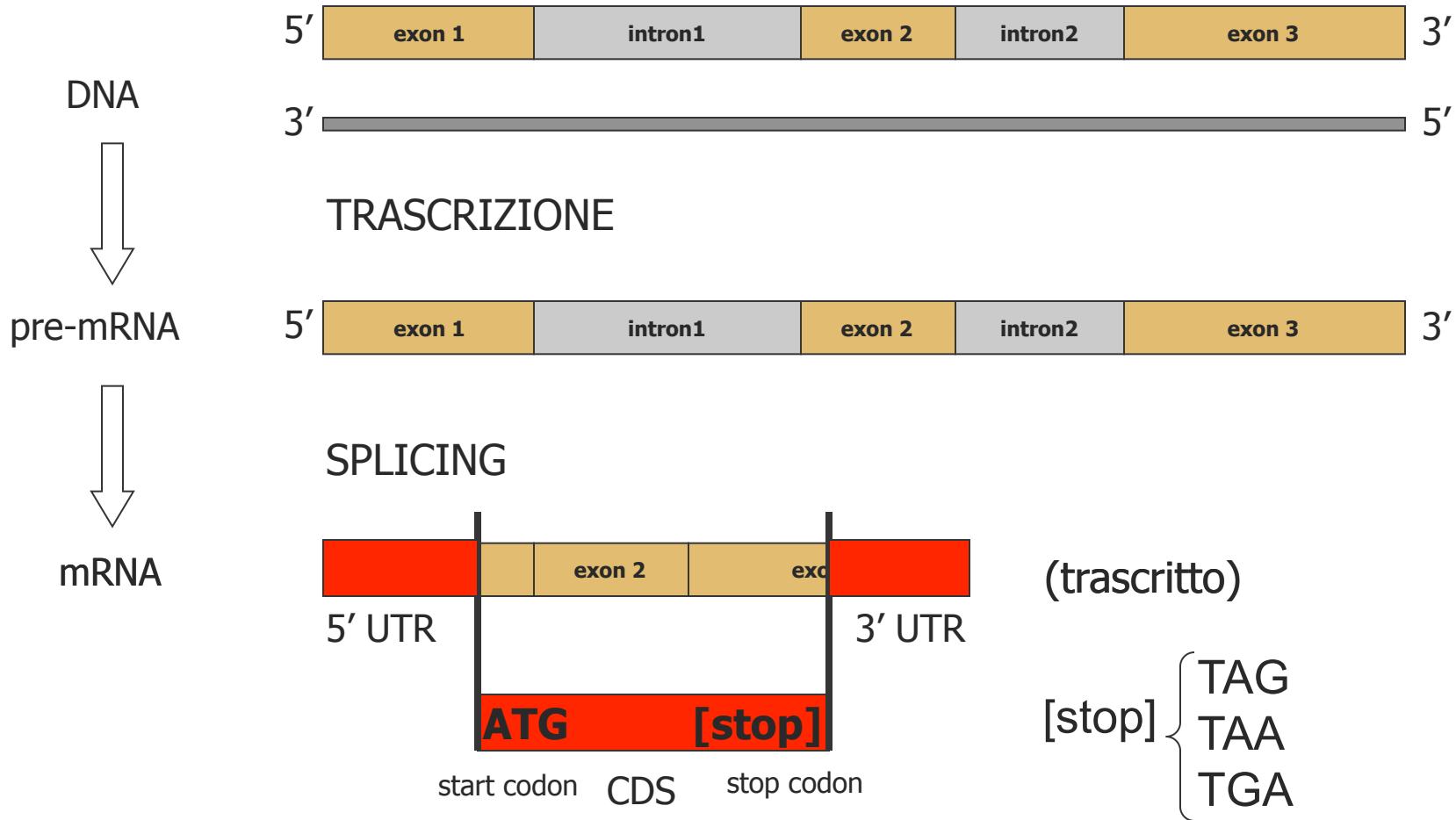
Espressione di un gene



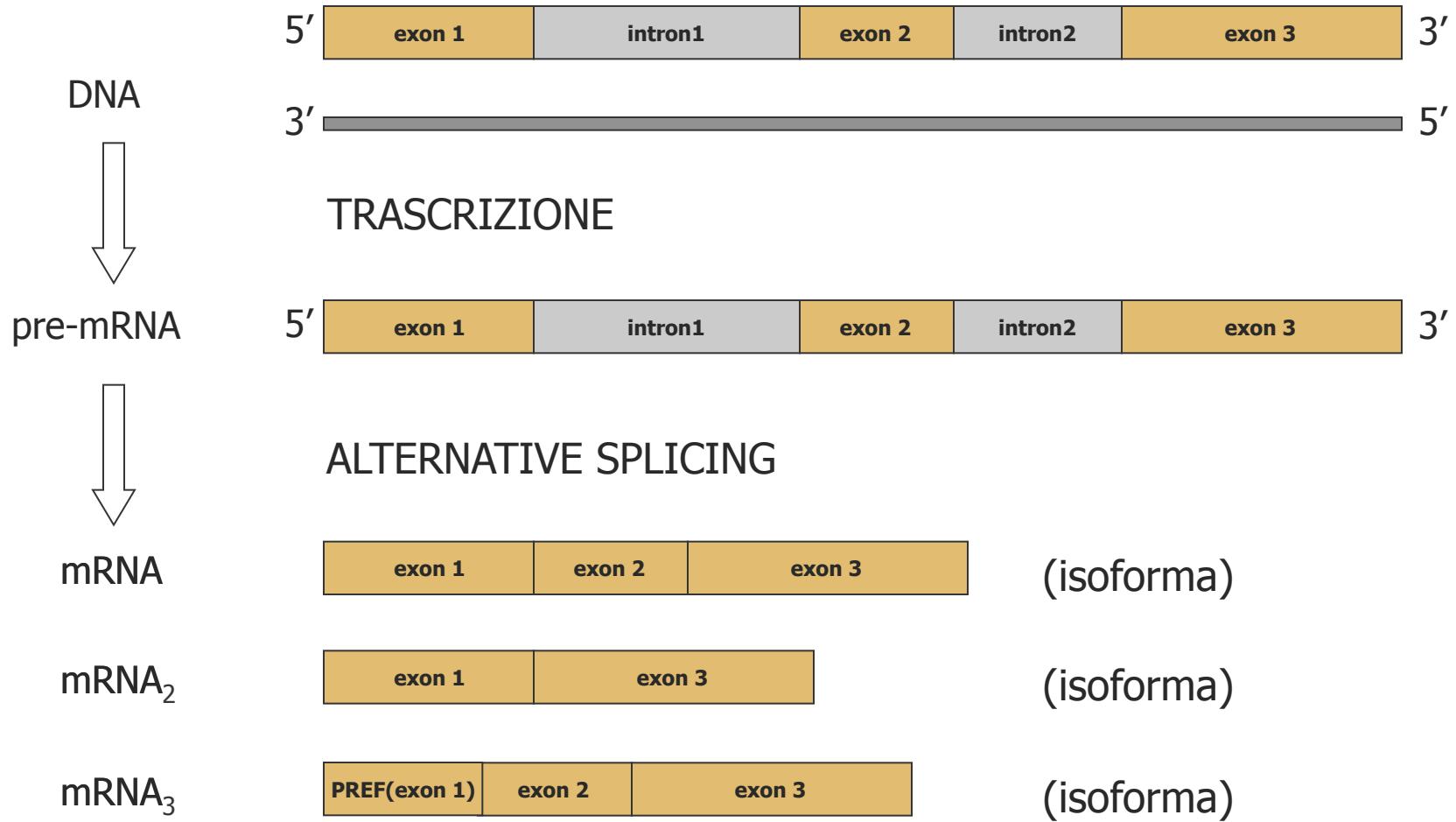
Espressione di un gene



Espressione di un gene



Espressione 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 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 exons
 - ✓ regioni non tradotte (UTR)
 - ✓ regioni non tradotte (UTR)
 - ✓ 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 exons
 - ✓ regioni non tradotte
 - ✓ 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 in esoni
- ✓ Attenzione! La sequenza presa come riferimento non deve essere necessariamente presa sulla catena di trascrizione del gene
- ✓ coding
- ✓ start e stop codon

[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)
- ✓ 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)
- ✓ ha estensione *.`gtf` oppure *.`gff`
- ✓ è composto da *record* (righe) di nove campi separati da tabulazione

[GTF (Gene Transfer Format)]

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

- ✓ un esone
- ✓ una CDS (o una sua parte)
- ✓ un 5' UTR (o una sua parte)
- ✓ un 3' UTR (o una sua parte)
- ✓ uno start codon (o una sua parte)
- ✓ uno stop codon (o una sua parte)

Feature relativa a un esone

Si supponga di avere un gene che esprime un trascritto composto da quattro esoni, e di conoscere la loro collocazione sulla genomica presa come riferimento (barra grigia).

5'

3'

mRNA

exon1

exon2

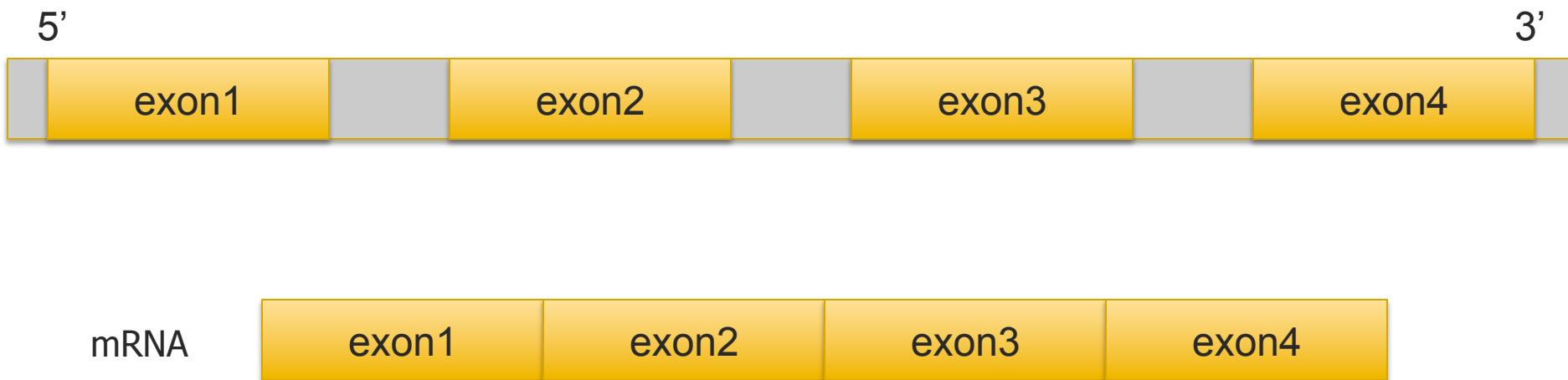
exon3

exon4

NB. La genomica (per ora) è sulla stessa catena di trascrizione del gene

Feature relativa a un esone

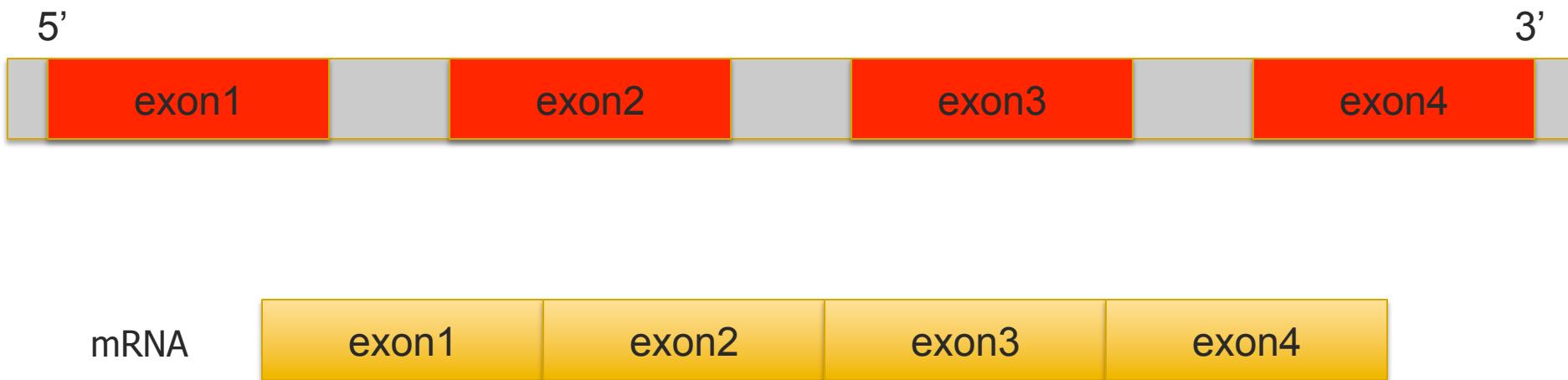
Si supponga di avere un gene che esprime un trascritto composto da quattro esoni, e di conoscere la loro collocazione sulla genomica presa come riferimento (barra grigia).



NB. La genomica (per ora) è sulla stessa catena di trascrizione del gene

[Feature relativa a un esone]

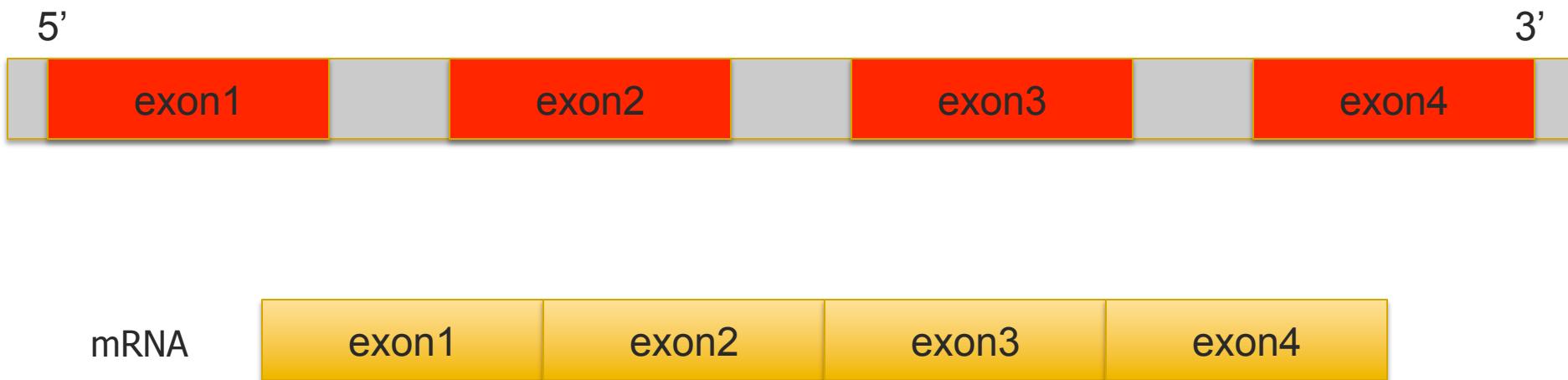
Si supponga di avere un gene che esprime un trascritto composto da quattro esoni, e di conoscere la loro collocazione sulla genomica presa come riferimento (barra grigia).



Le sottostringhe rosse sono le quattro *features* che corrispondono ai quattro esoni del trascritto.

[Feature relativa a un esone]

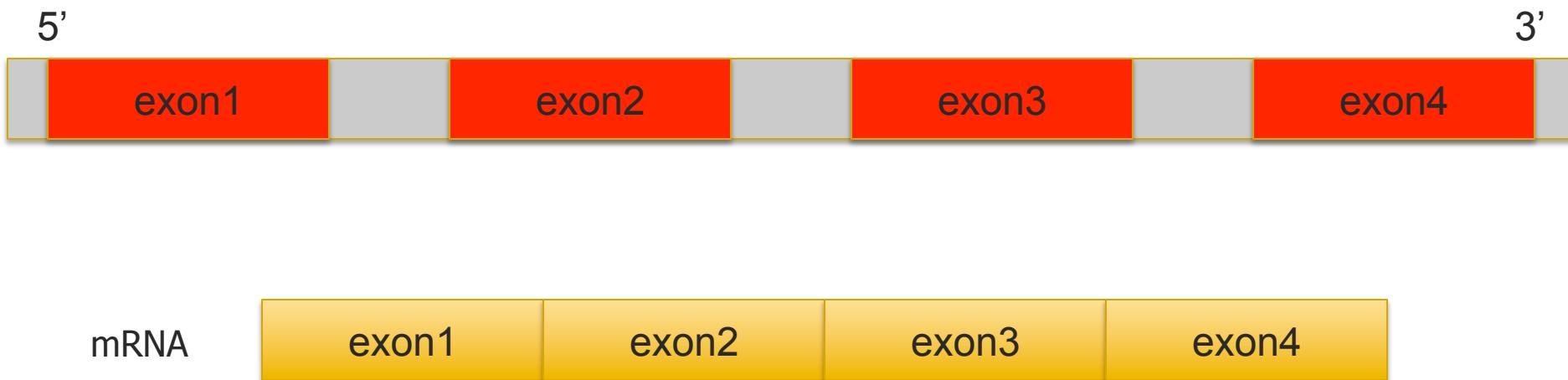
Si supponga di avere un gene che esprime un trascritto composto da quattro esoni, e di conoscere la loro collocazione sulla genomica presa come riferimento (barra grigia).



Ad ogni esone del trascritto corrisponde una feature sulla sequenza di riferimento.

[Feature relativa a un esone]

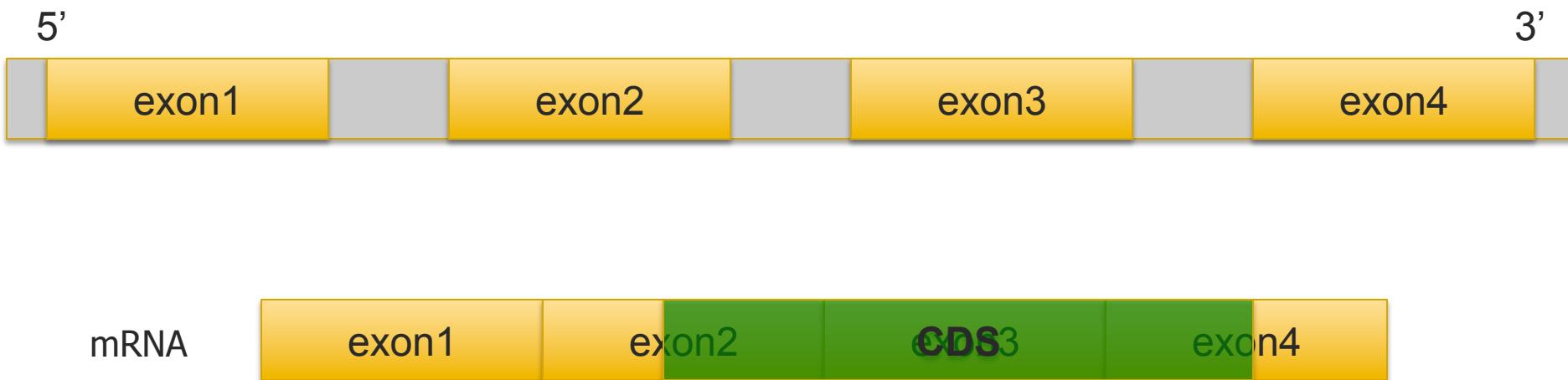
Si supponga di avere un gene che esprime un trascritto composto da quattro esoni, e di conoscere la loro collocazione sulla genomica presa come riferimento (barra grigia).



Ad ogni esone del trascritto corrisponde un record in formato GTF.

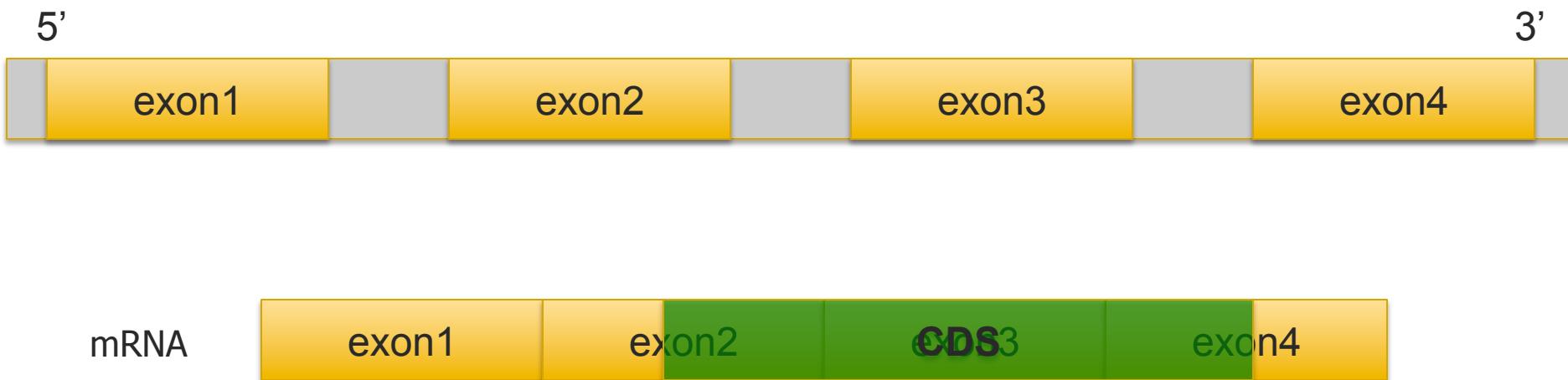
[*Features relative a una CDS*]

Si supponga di conoscere la coding sequence (CDS) sul trascritto (evidenziata in verde).



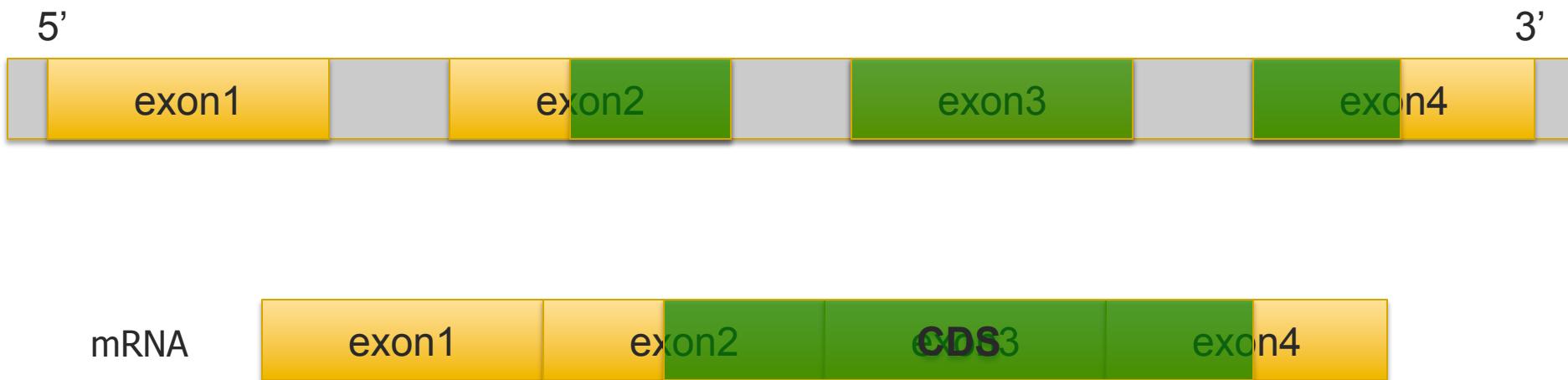
[*Features relative a una CDS*]

Si riporti la CDS del trascritto sulla sequenza genomica di riferimento.



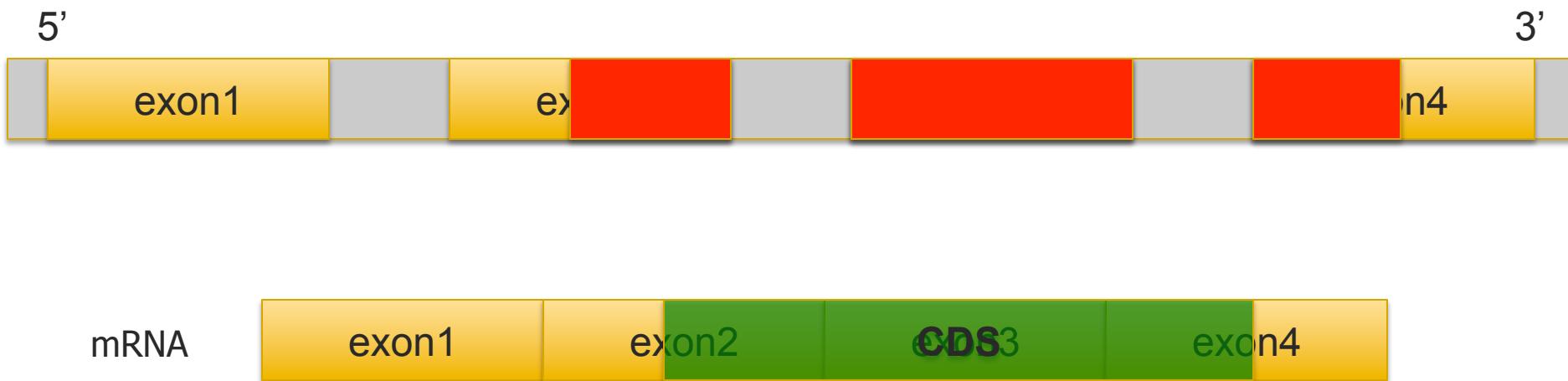
[*Features relative a una CDS*]

Si riporti la CDS del trascritto sulla sequenza genomica di riferimento.



[Features relative a una CDS]

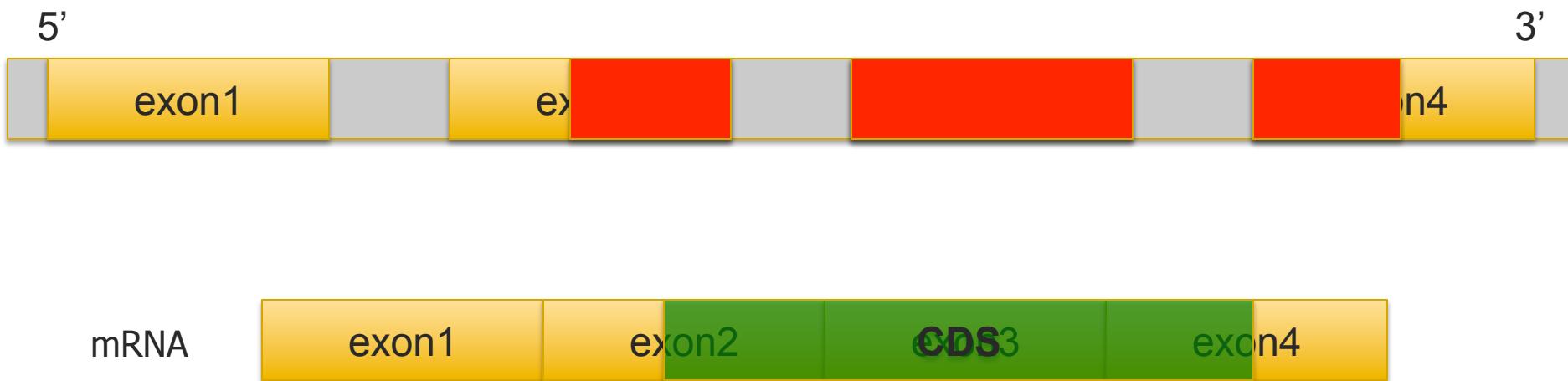
Si riporti la CDS del trascritto sulla sequenza genomica di riferimento.



Le sottostringhe rosse sono le tre *features* sulla genomica corrispondenti alla CDS sul trascritto.

[Features relative a una CDS]

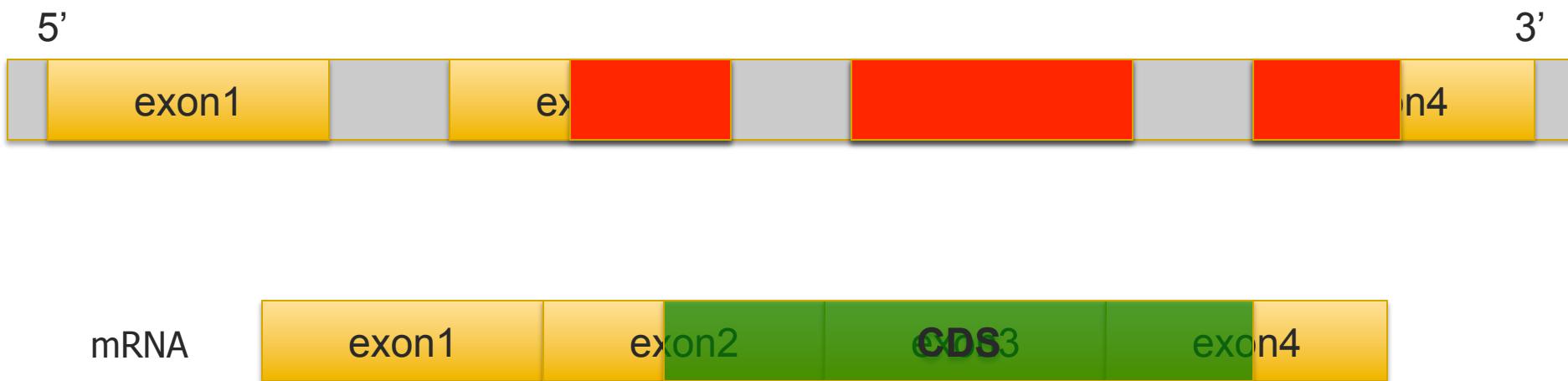
Si riporti la CDS del trascritto sulla sequenza genomica di riferimento.



Ad una CDS di un trascritto corrispondono una o più *features* sulla sequenza di riferimento.

[Features relative a una CDS]

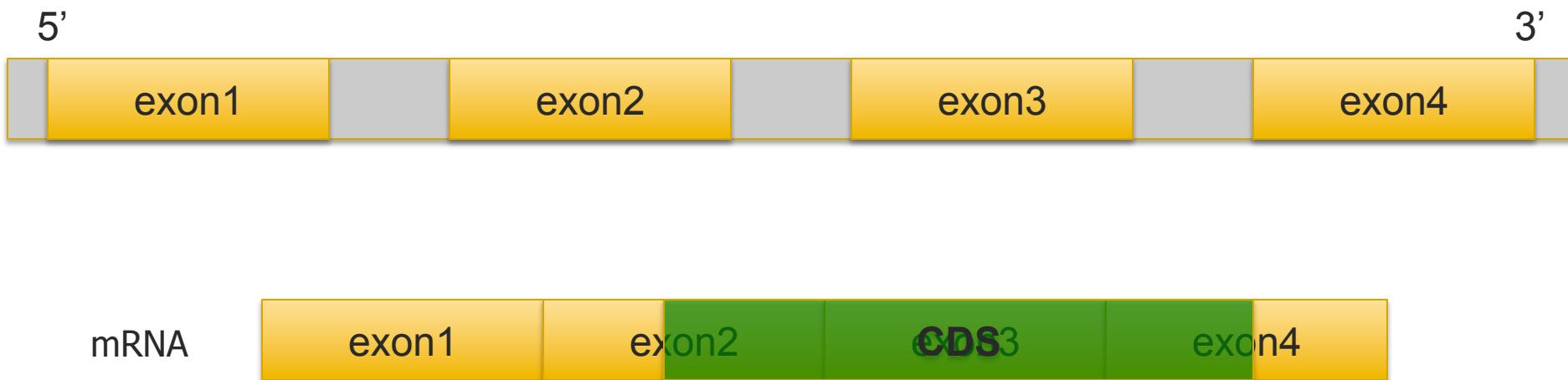
Si riporti la CDS del trascritto sulla sequenza genomica di riferimento.



Ad una CDS di un trascritto corrispondono uno o più record in formato GTF.

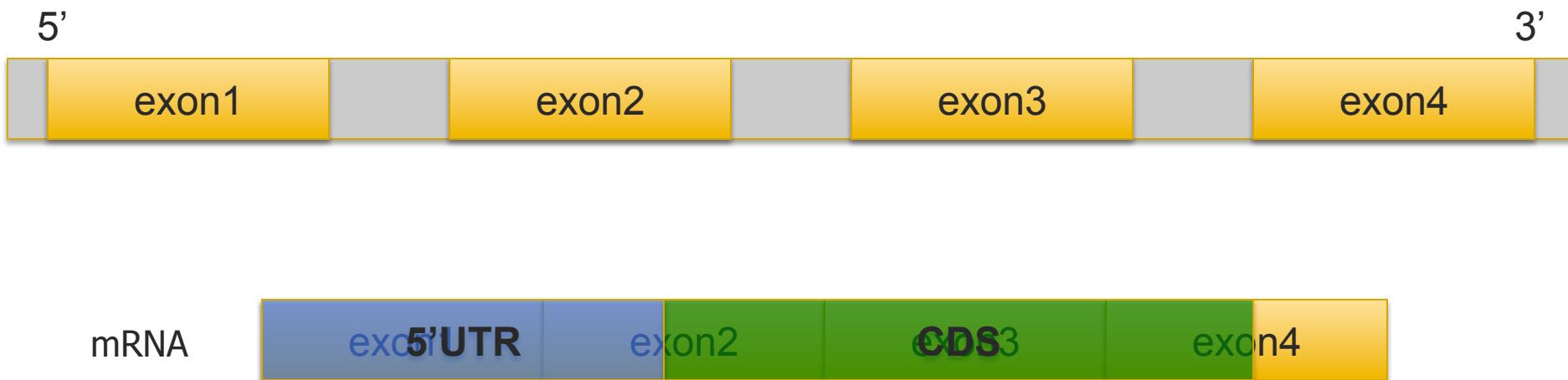
[Features relative a un 5' UTR]

Il 5' UTR è il prefisso di trascritto che precede la CDS



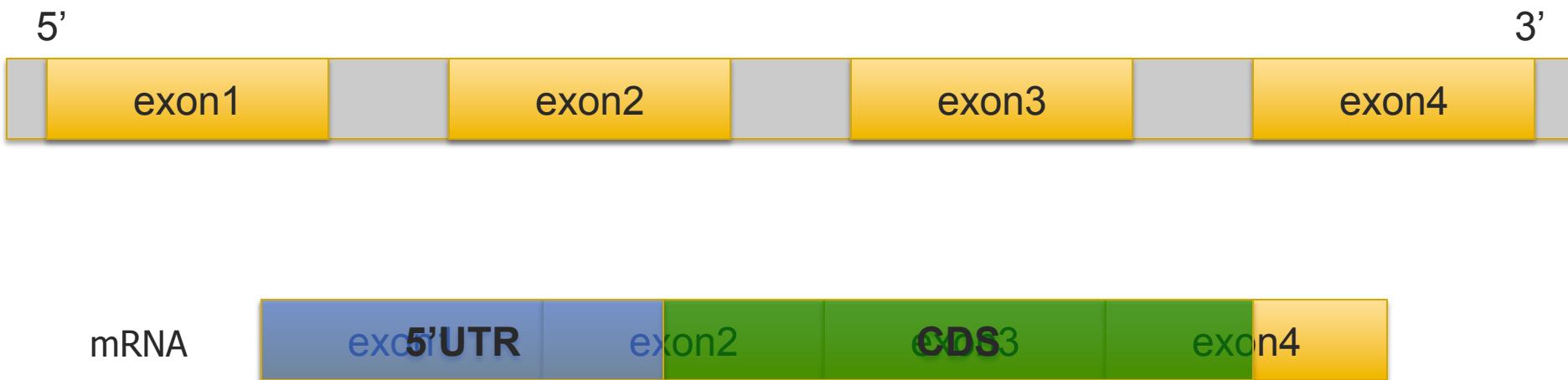
[Features relative a un 5' UTR]

Il 5' UTR è il prefisso di trascritto che precede la CDS



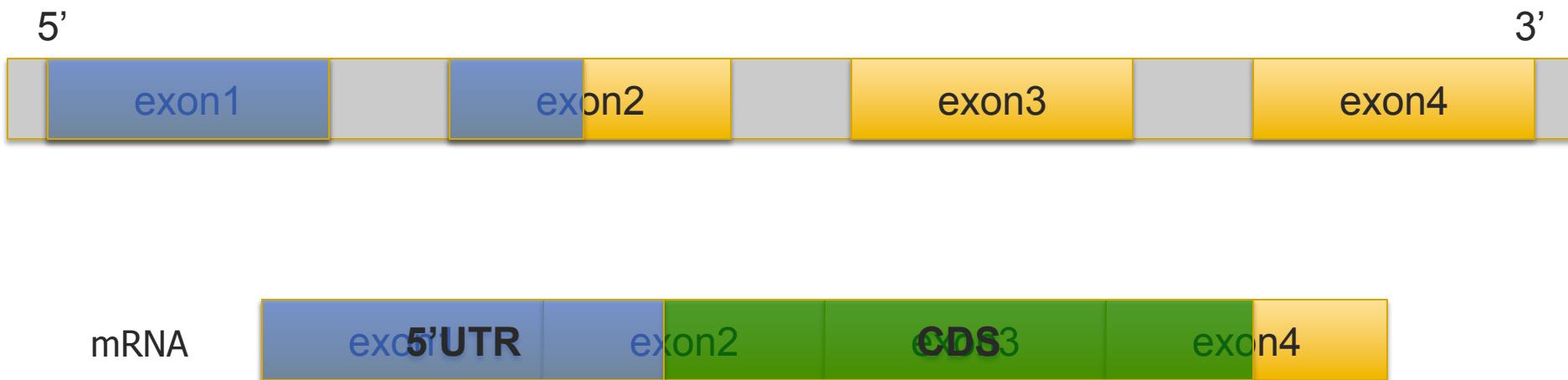
[Features relative a un 5' UTR]

Si riporti il 5' UTR del trascritto sulla sequenza genomica di riferimento.



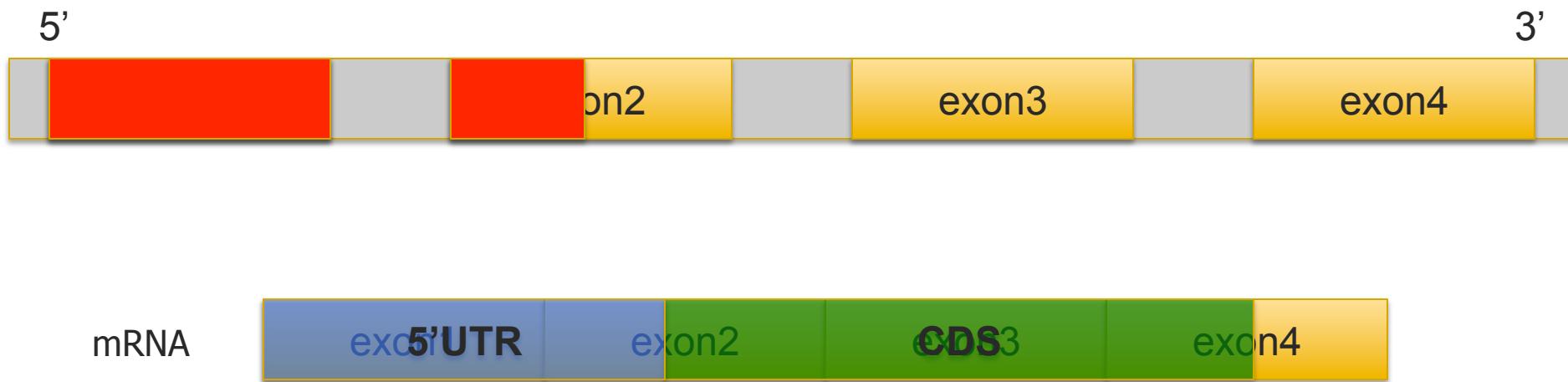
[Features relative a un 5' UTR]

Si riporti il 5' UTR del trascritto sulla sequenza genomica di riferimento.



[Features relative a un 5' UTR]

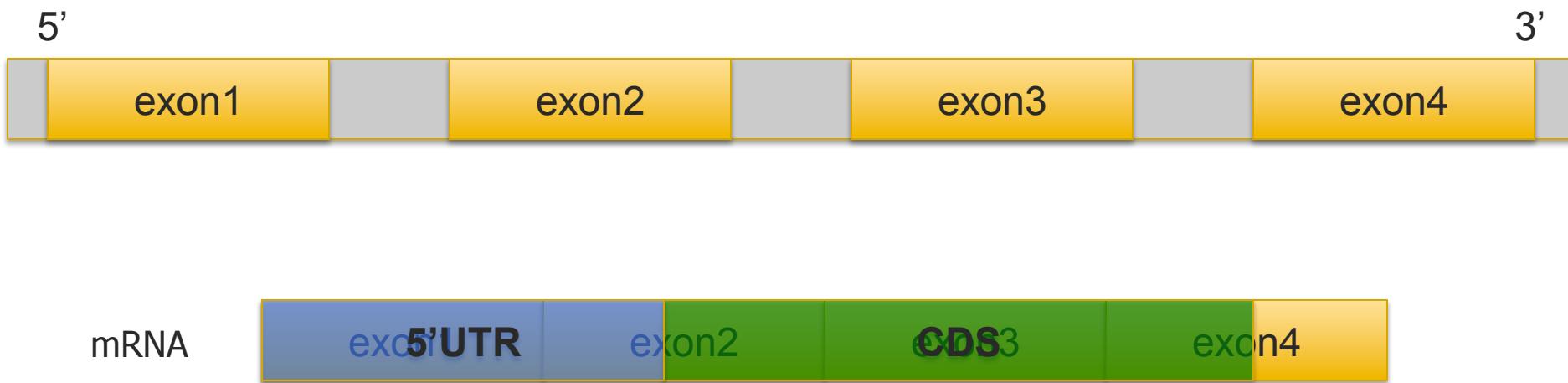
Si riporti il 5' UTR del trascritto sulla sequenza genomica di riferimento.



Le sottostringhe rosse sono le due *features* sulla genomica corrispondenti al 5' UTR sul trascritto.

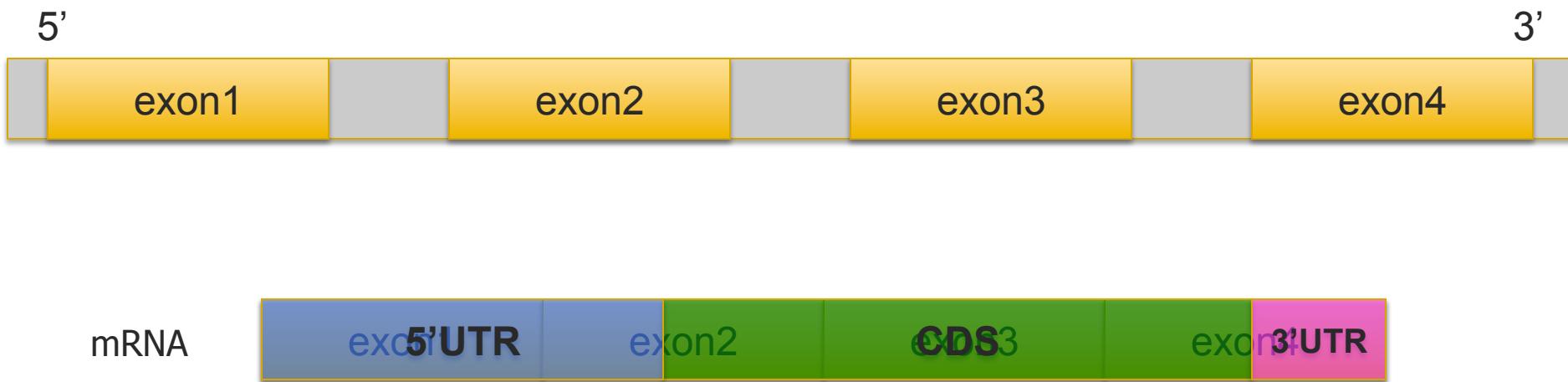
[Features relative a un 3' UTR]

Il 3' UTR è il suffisso di trascritto che segue la CDS



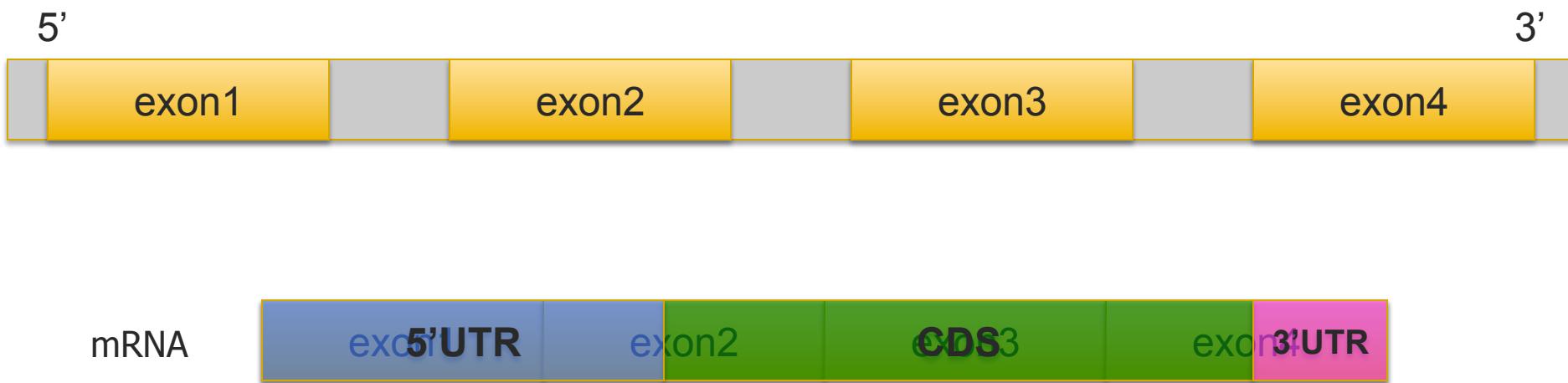
[Features relative a un 3' UTR]

Il 3' UTR è il suffisso di trascritto che segue la CDS



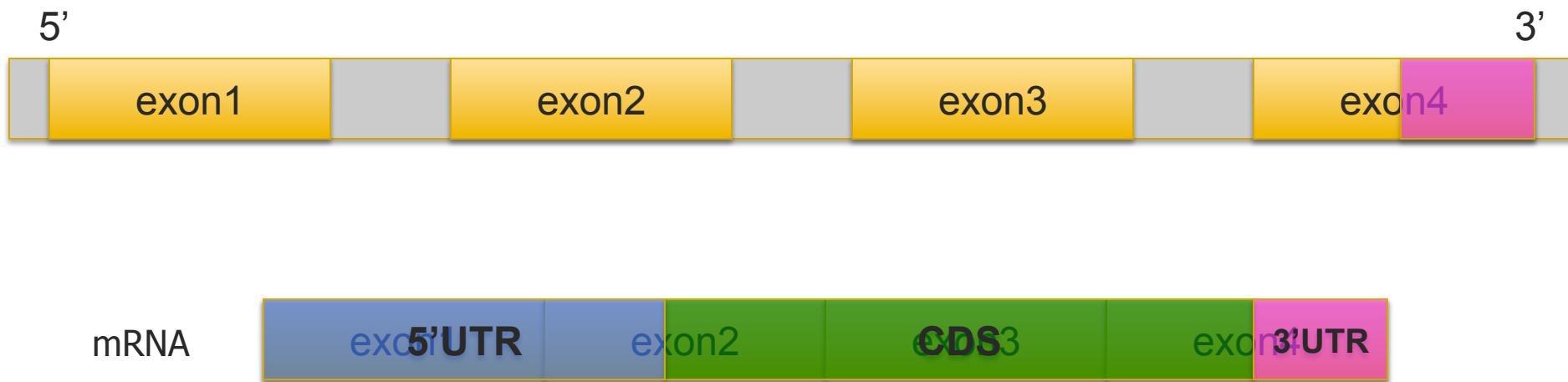
[Features relative a un 3' UTR]

Si riporti il 3' UTR del trascritto sulla sequenza genomica di riferimento.



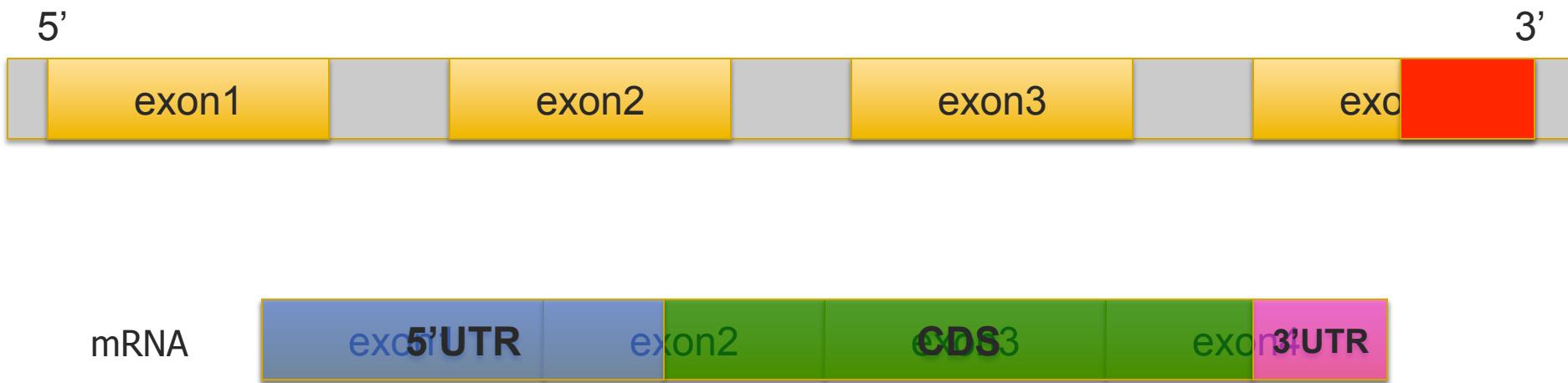
[Features relative a un 3' UTR]

Si riporti il 3' UTR del trascritto sulla sequenza genomica di riferimento.



[Features relative a un 3' UTR]

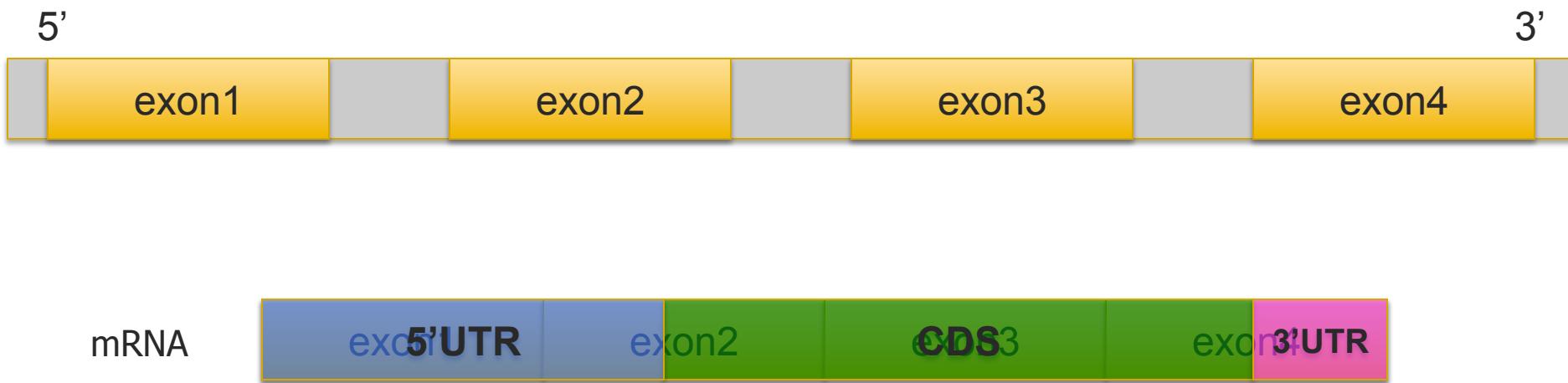
Si riporti il 3' UTR del trascritto sulla sequenza genomica di riferimento.



La sottostringa rossa è la *feature* sulla genomica corrispondente al 3' UTR sul trascritto.

[Features relative a un 3' UTR]

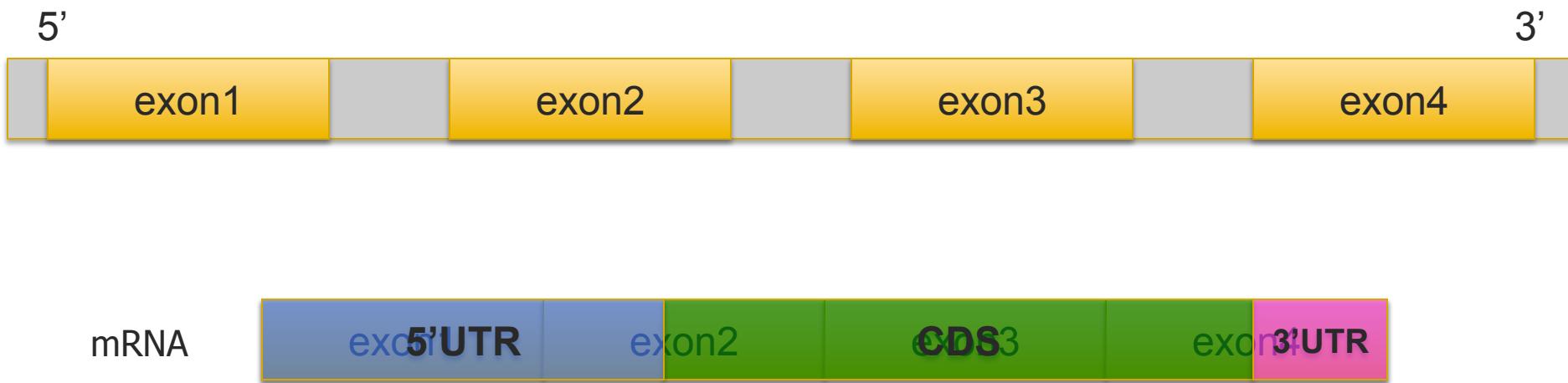
Si riporti il 3' UTR del trascritto sulla sequenza genomica di riferimento.



Ad un 5'UTR/3'UTR di un trascritto corrispondono una o più *features* sulla sequenza di riferimento.

[Features relative a un 3' UTR]

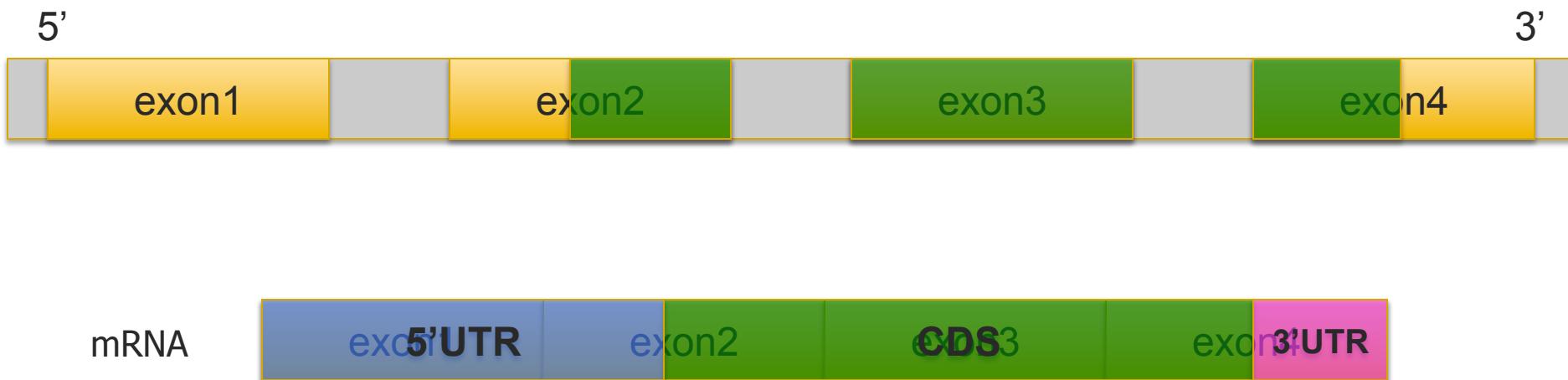
Si riporti il 3' UTR del trascritto sulla sequenza genomica di riferimento.



Ad un 5'UTR/3'UTR di un trascritto corrispondono uno o più record in formato GTF.

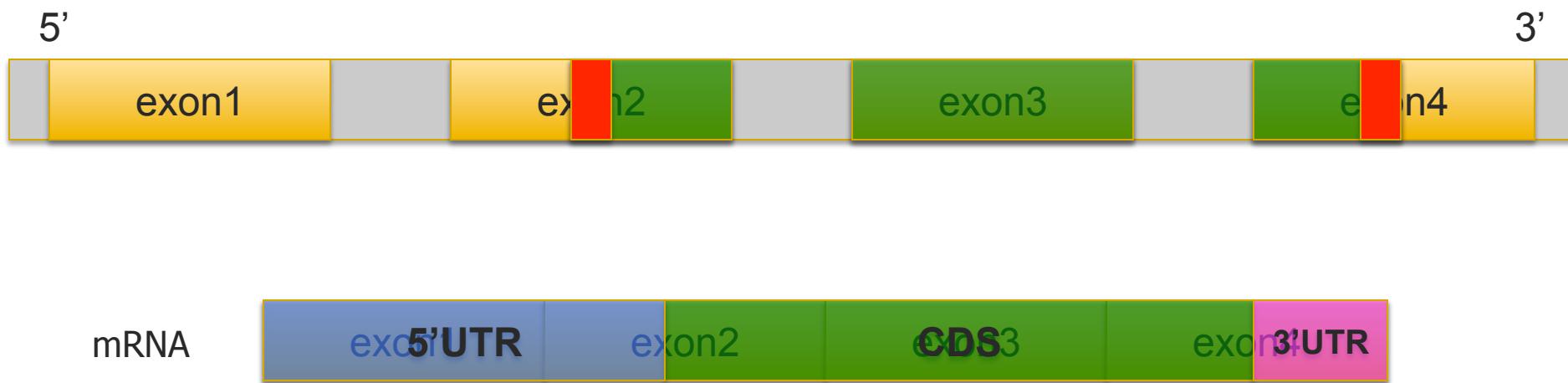
[Features relative a start/stop]

Anche allo start e allo stop codon corrispondono delle *features* sulla sequenza di riferimento.



[Features relative a start/stop]

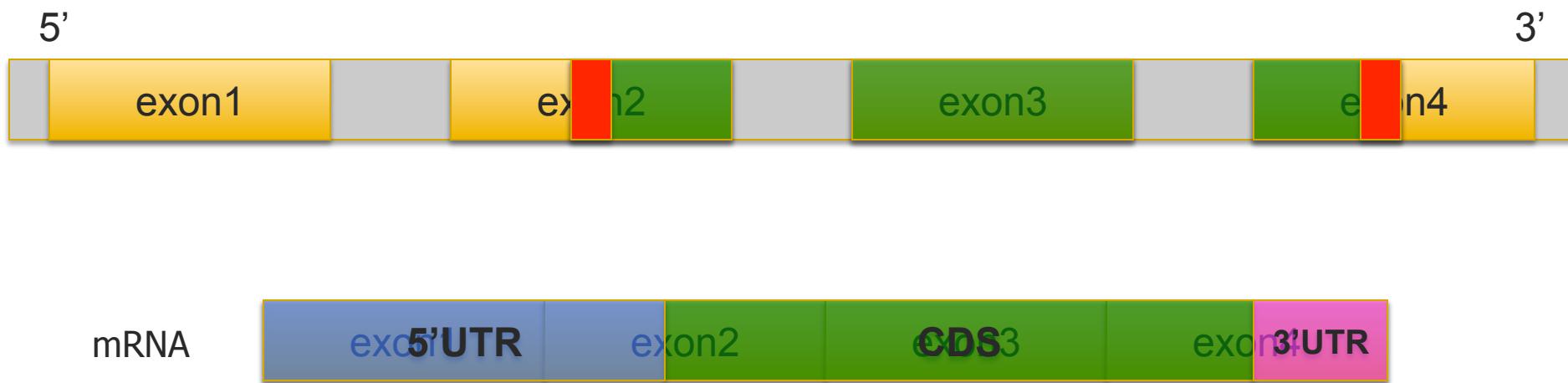
Anche allo start e allo stop codon corrispondono delle *features* sulla sequenza di riferimento.



Le due sottostringhe rosse (lunghe tre basi) sono le due *features* che descrivono rispettivamente lo start e lo stop codon della CDS.

[Features relative a start/stop]

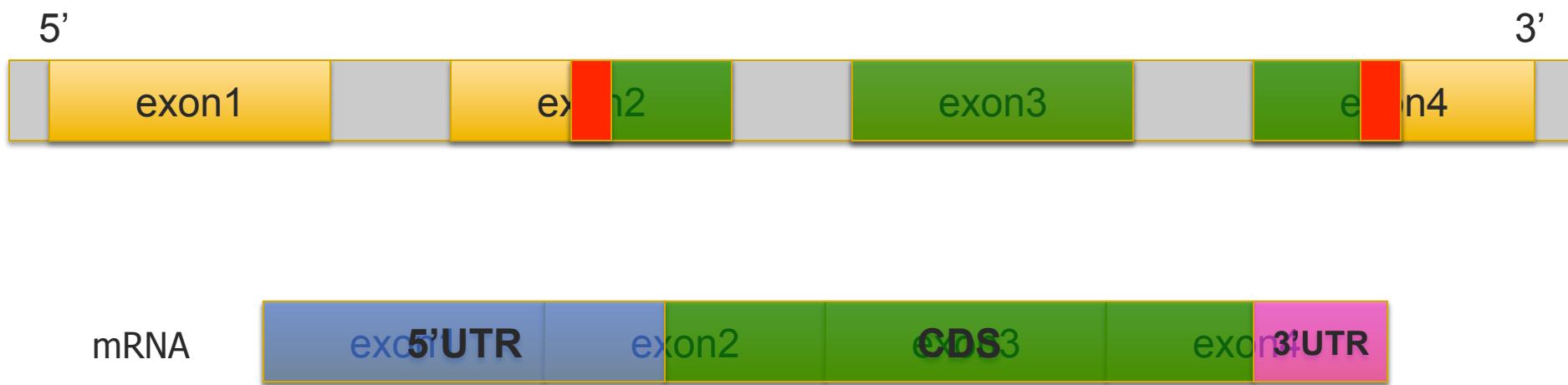
Anche allo start e allo stop codon corrispondono delle *features* sulla sequenza di riferimento.



Ad uno start/stop codon di un trascritto corrispondono una o più *features* sulla sequenza di riferimento.

[Features relative a start/stop]

Anche allo start e allo stop codon corrispondono delle *features* sulla sequenza di riferimento.



Ad uno start/stop codon di un trascritto corrispondono uno o più record in formato GTF.

I campi di un *record* GTF

In generale, ogni *record* di un file GTF descrive una *feature* sulla sequenza genomica di riferimento, ed è composto dai seguenti nove campi:

I campi di un *record* GTF

In generale, ogni *record* di un file GTF descrive una *feature* sulla sequenza genomica di riferimento, ed è composto dai seguenti nove campi:

- ✓ identificatore della genomica di riferimento

I campi di un *record* GTF

In generale, ogni *record* di un file GTF descrive una *feature* sulla sequenza genomica di riferimento, ed è composto dai seguenti nove campi:

- ✓ identificatore della genomica di riferimento
- ✓ la sorgente che ha prodotto l'annotazione (ad esempio un *software*)

I campi di un *record* GTF

In generale, ogni *record* di un file GTF descrive una *feature* sulla sequenza genomica di riferimento, ed 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"}

I campi di un *record* GTF

In generale, ogni *record* di un file GTF descrive una *feature* sulla sequenza genomica di riferimento, ed 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*
- ✓ la posizione di inizio della *feature* sulla genomica di riferimento
- ✓ la posizione di fine della *feature* sulla genomica di riferimento

I campi di un *record* GTF

In generale, ogni *record* di un file GTF descrive una *feature* sulla sequenza genomica di riferimento, ed 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*
- ✓ 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)

I campi di un *record* GTF

In generale, ogni *record* di un file GTF descrive una *feature* sulla sequenza genomica di riferimento, ed 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*
- ✓ 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: {+, -}

+ , se la genomica di riferimento è sulla catena di trascrizione del gene
- , se la genomica di riferimento è sulla catena opposta

I campi di un *record* GTF

In generale, ogni *record* di un file GTF descrive una *feature* sulla sequenza genomica di riferimento, ed 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*
- ✓ 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*
- ✓ il *frame* ('.', se alla *feature* non è associato un *frame*)

Valori possibili: {0, 1, 2}

Solo *features* "CDS", "start_codon" e "stop_codon" hanno un valore del *frame*

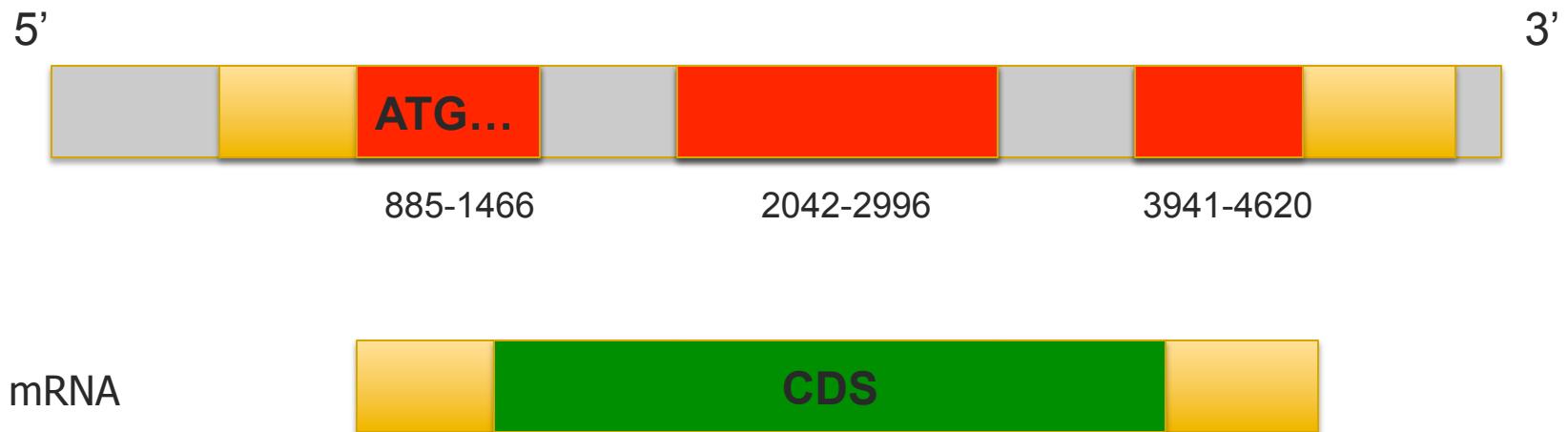
I campi di un *record* GTF

In generale, ogni *record* di un file GTF descrive una *feature* sulla sequenza genomica di riferimento, ed 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*
- ✓ 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*
- ✓ il *frame* ('.', se alla *feature* non è associato un *frame*)
- ✓ il campo degli attributi nella forma:
 - ✓ <attribute_name1> <value1>; <attribute_name2> <value2>; ...
 - ✓ Attributi obbligatori sono: "gene_id" e "transcript_id"

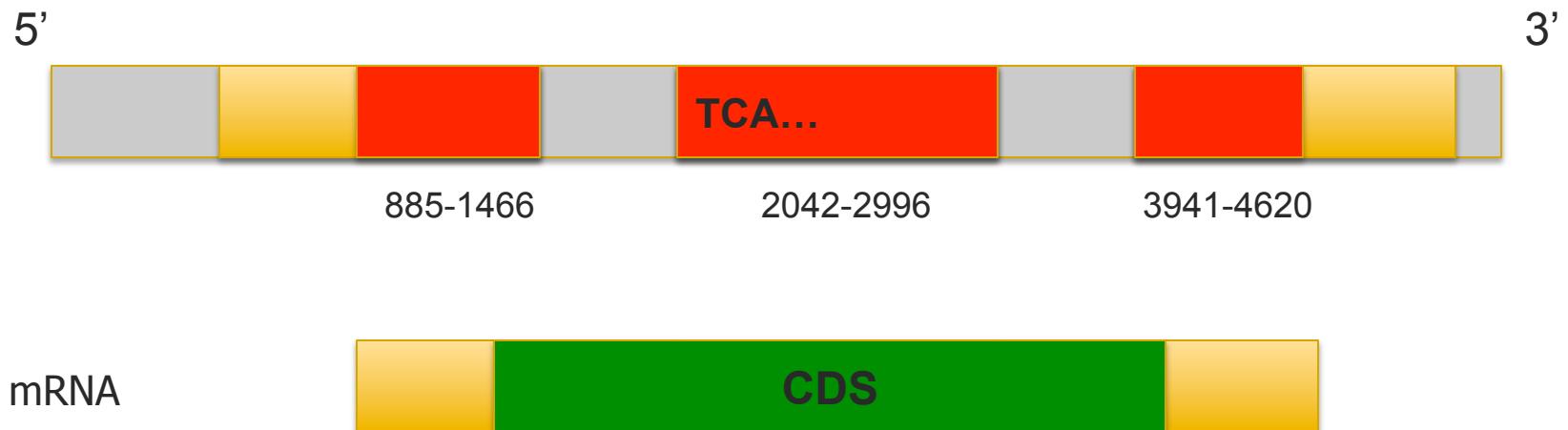
Il frame

Osservazione “banale”: i primi tre simboli della prima *feature* “cds” 885-1466 corrispondono al codone di inizio “ATG” della CDS.



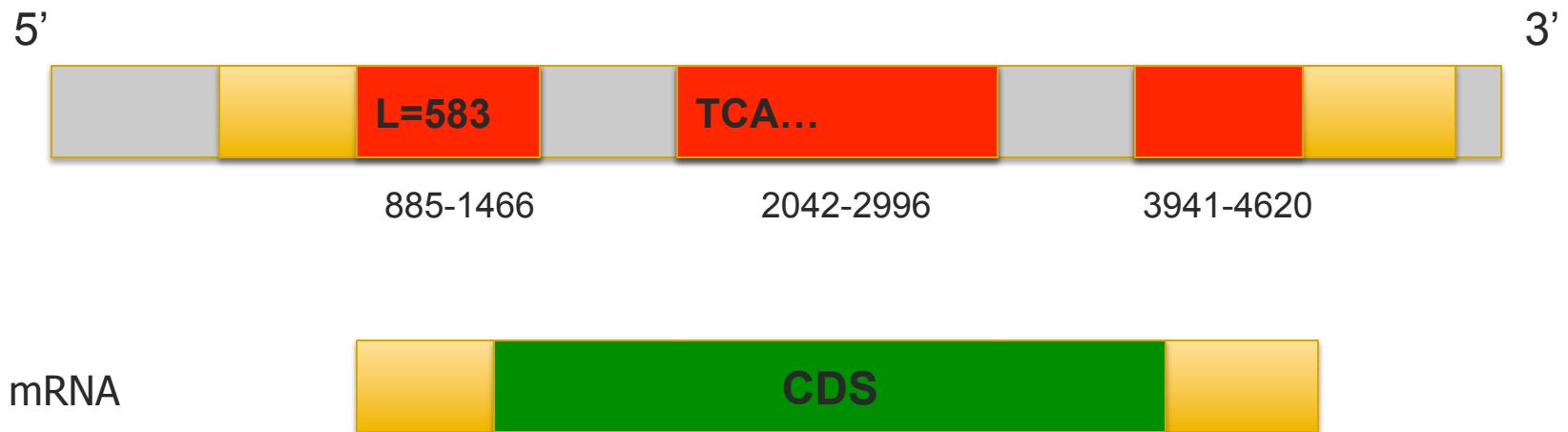
Il frame

Domanda1: i primi tre simboli della seconda *feature* 2042-2996 sono un codone della CDS? Supponiamo che siano “TCA”



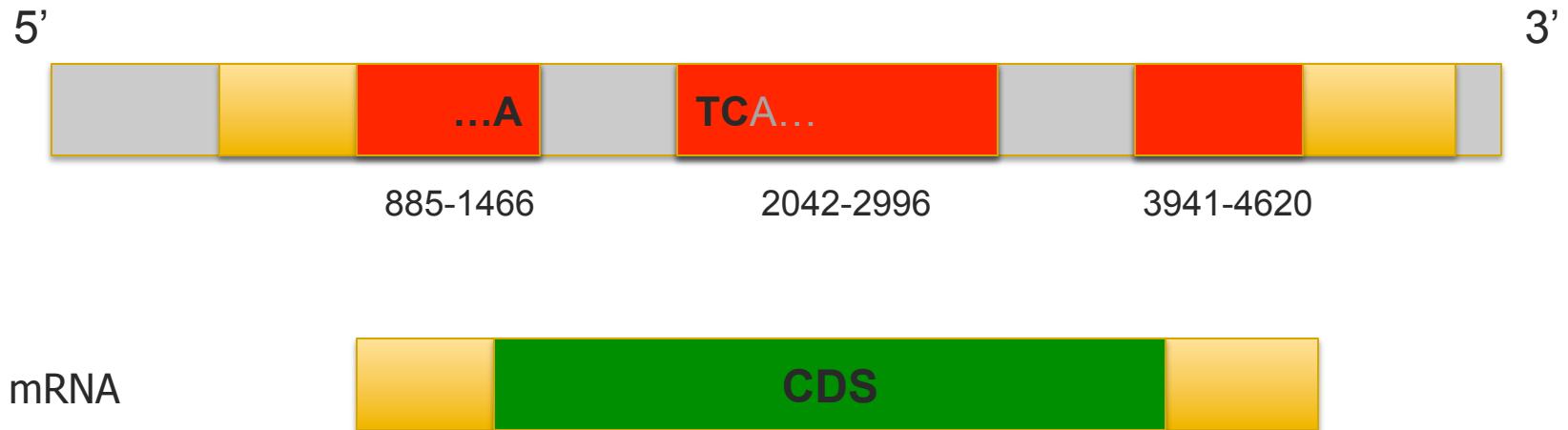
Il frame

Risposta1: no, perché la lunghezza della prima *feature* (583 bp) non è un multiplo di tre!



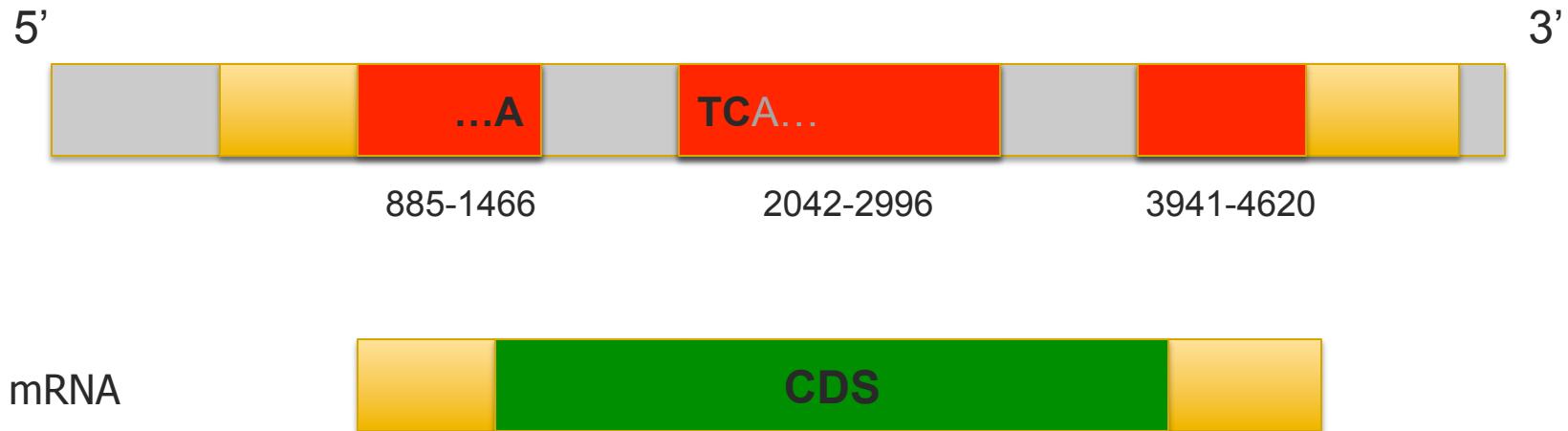
Il frame

Dal momento che il resto della divisione intera tra 583 e 3 è pari a 1, l'ultima base della prima *feature* 885-1466 (supponiamo “A”) risulta essere la prima base del codone “ATC” che ha le ultime due basi T e C all’inizio della seconda feature 2042-2996



Il frame

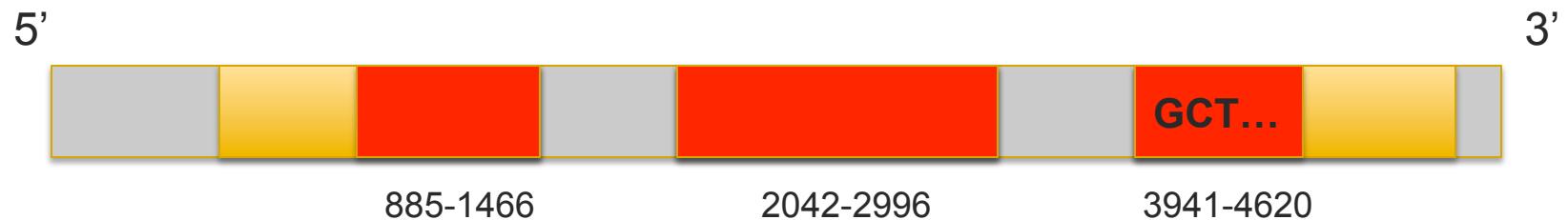
Dal momento che il resto della divisione intera tra 583 e 3 è pari a 1, l'ultima base della prima *feature* 885-1466 (supponiamo “A”) risulta essere la prima base del codone “ATC” che ha le ultime due basi T e C all’inizio della seconda feature 2042-2996



Il fatto che le prime due basi della seconda *feature* 2042-2996 sono le ultime due basi di un codone viene espresso dicendo che tale *feature* ha *frame* pari a 1

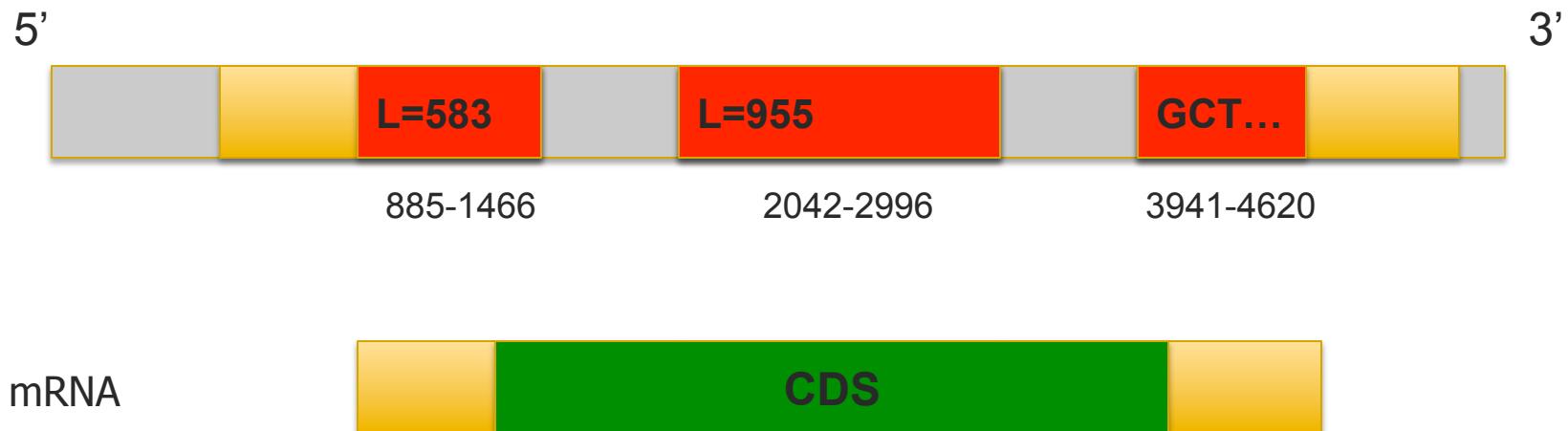
Il frame

Domanda2: i primi tre simboli della terza *feature* 3941-4620 sono un codone della CDS? Supponiamo che siano “GCT”



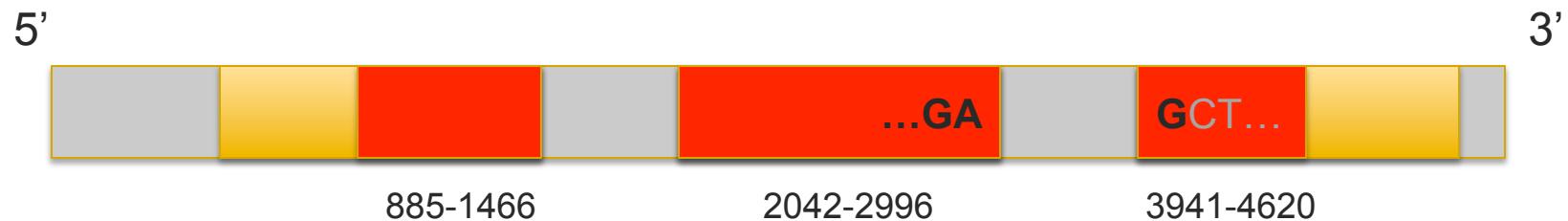
Il frame

Risposta: no, perché la somma delle lunghezze delle *features* precedenti ($583 \text{ bp} + 955 \text{ bp} = 1538 \text{ bp}$) non è un multiplo di tre!



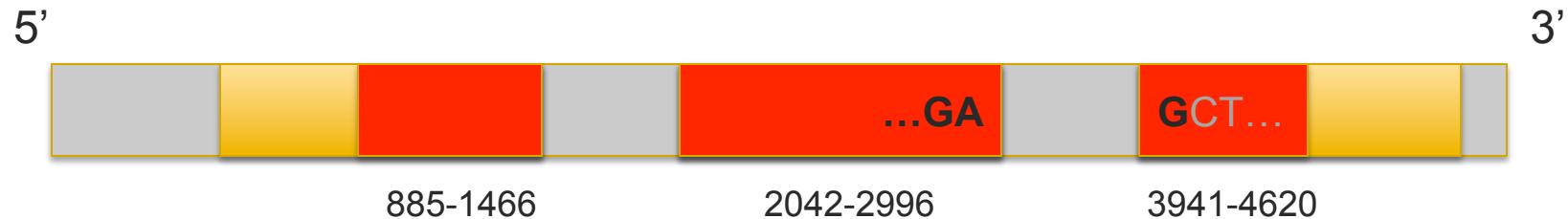
Il frame

Dal momento che il resto della divisione intera tra 1538 e 3 è pari a 2, le ultime due basi della seconda *feature* 2042-2996 (supponiamo “GA”) risultano essere le prime due basi del codone “GAG” che ha l’ultima base G all’inizio della terza *feature* 3941-4620



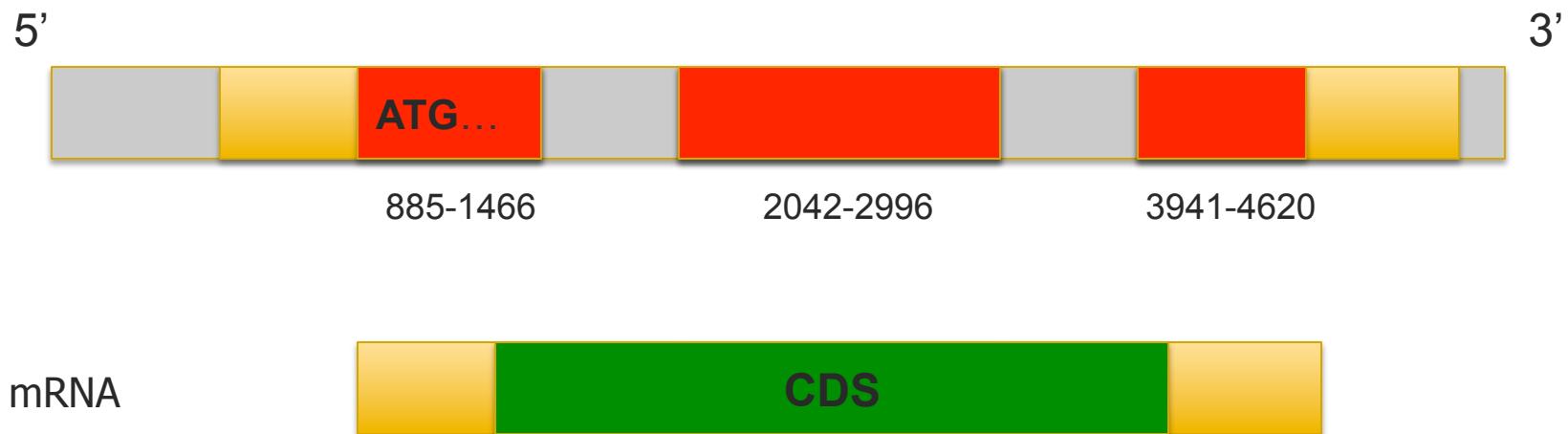
Il frame

Dal momento che il resto della divisione intera tra 1538 e 3 è pari a 2, le ultime due basi della seconda *feature* 2042-2996 (supponiamo “GA”) risultano essere le prime due basi del codone “GAG” che ha l’ultima base G all’inizio della terza *feature* 3941-4620



Il fatto che la prima base della terza *feature* 3941-4620 è l’ultima base di un codone viene espresso dicendo che tale *feature* ha *frame* pari a 2

Il frame



Il fatto che la prima base di un *feature* “CDS” è la prima base di un codone (come per la prima *feature* 885-1466) viene espresso dicendo che tale *feature* ha *frame* pari a 0

[Il frame]

Il *frame* di una *feature* F, che corrisponde a una CDS, è calcolato nel seguente modo:

- ✓ si determina la lunghezza totale L delle *features* precedenti a F
- ✓ si calcola il resto della divisione intera di L e 3 (tale resto può essere pari a 0, oppure a 1, oppure a 2)
- ✓ Il frame di F è posto uguale a tale resto

[Il frame]

Il *frame* di una *feature* F, che corrisponde a una CDS, è calcolato nel seguente modo:

- ✓ si determina la lunghezza totale L delle *features* precedenti a F
- ✓ si calcola il resto della divisione intera di L e 3 (tale resto può essere pari a 0, oppure a 1, oppure a 2)
- ✓ Il frame di F è posto uguale a tale resto

Un *frame* pari a 0 significa che la prima base di F coincide con la prima base di un codone della CDS.

[Il frame]

Il *frame* di una *feature* F, che corrisponde a una CDS, è calcolato nel seguente modo:

- ✓ si determina la lunghezza totale L delle *features* precedenti a F
- ✓ si calcola il resto della divisione intera di L e 3 (tale resto può essere pari a 0, oppure a 1, oppure a 2)
- ✓ Il frame di F è posto uguale a tale resto

Un *frame* pari a 1 significa che le prime due basi di F coincidono con le ultime due basi di un codone della CDS. La prima base del codone coincide con l'ultima base della *feature* che precede F.

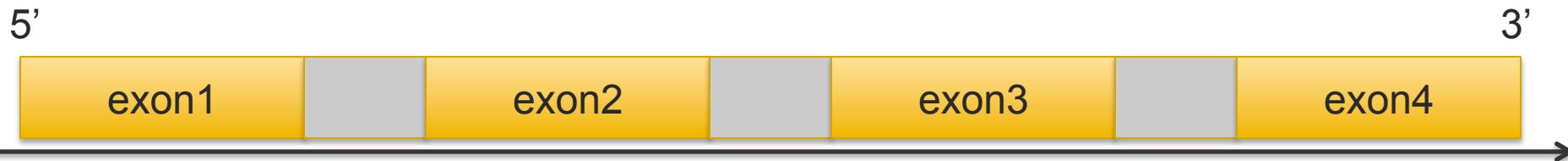
[Il frame]

Il *frame* di una *feature* F, che corrisponde a una CDS, è calcolato nel seguente modo:

- ✓ si determina la lunghezza totale L delle *features* precedenti a F
- ✓ si calcola il resto della divisione intera di L e 3 (tale resto può essere pari a 0, oppure a 1, oppure a 2)
- ✓ Il frame di F è posto uguale a tale resto

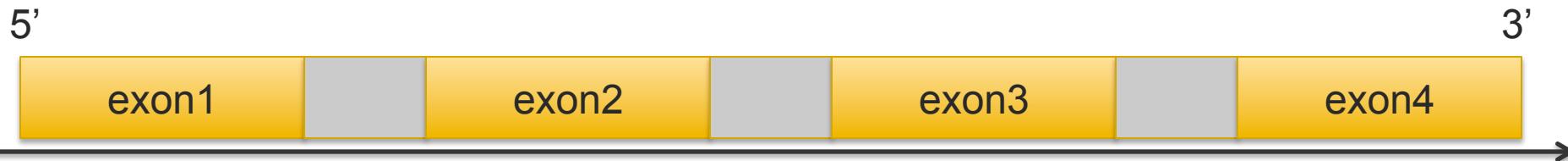
Un *frame* pari a 2 significa che la prima base di F coincide con l'ultima base di un codone della CDS. Le prime due basi del codone coincidono con le ultime due basi della *feature* che precede F.

[Lo strand]

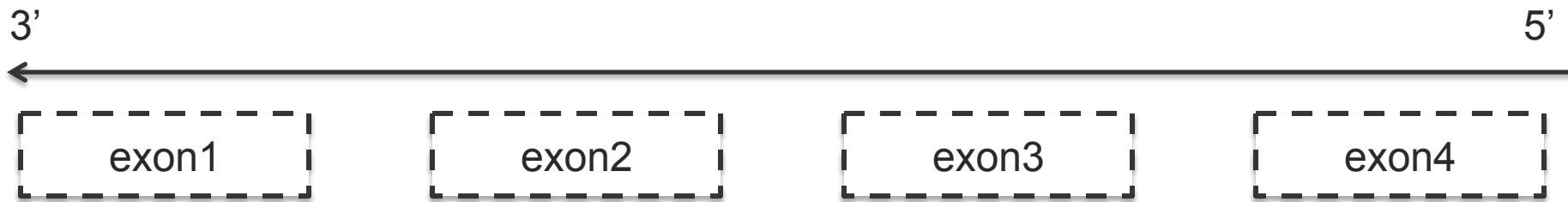


Annotare un gene su di una sequenza di riferimento presa sulla sua catena di trascrizione implica uno *strand +*

[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]

Uno *strand* – implica che, per ottenere la sequenza di una determinata *feature*, si deve:

- ① estrarre la sottostringa della genomica di riferimento che corrisponde alla *feature*
- ② eseguire un'operazione di reverse&complement della sottostringa

[Esercizio]

Scrivere un programma che prenda in input un file in formato GTF e il file FASTA della sequenza di riferimento relativa al GTF, e produca in output un file FASTA contenente le sequenze di tutti i trascritti e di tutte le CDS dei geni annotati nel file GTF.

Suggerimenti:

- il metodo `split` della classe `String` divide la stringa usando il separatore passato come parametro, e restituisce un array contenente le singole parti. Se il parametro non viene specificato, allora viene usato (come separatore) lo spazio.