Luigi Avenoso 844787

Validatore file GTF

La validazione del file è stata sviluppata solamente a livello sintattico, così come le violazioni in esso contenute sono solo di tipo sintattico.

Il controllo del file non si arresta al primo errore trovato ma continua fino alla fine riportando tutti gli errori riscontrati.

Ogni riga deve contenere i nove campi elencati sotto (ad esclusione dei commenti che non sono obbligatori). Ogni campo deve essere separato dal successivo da un TAB. Un numero differente da nove è considerato errore.

Il campo **seqname** deve essere unico per tutti i record relativi a un dato gene. La validazione di questo campo si basa su una statistica, che considera come valore corretto quello più frequente per il relativo gene. Se la condizione non è rispettata viene considerato errore.

Il campo **source** deve essere univoco per tutto il file altrimenti è considerato errore.

Il campo **feature** deve essere identico ad uno tra i seguenti valori: "CDS","start_codon","stop_codon","5UTR","3UTR","inter","inter_CNS","intron_CNS","exon". Un qualsiasi valore diverso da essi viene considerato errore.

I campi **start** e **end** devono essere valori interi postivi e start deve essere minore o uguale a end. Lo 0 è escluso perché la numerazione parte da 1. Se le condizioni non sono rispettate è considerato errore.

Il campo **score** può essere un numero in virgola mobile o un intero e, non essendo necessario, può essere sostituito con un punto. Qualsiasi valore diverso dai precedenti è da considerarsi errore.

Il campo **strand** può assumere solamente i valori "+" o "-", qualsiasi altro valore è considerato un errore. La verifica che tutto il gene sia su un solo strand, positivo o negativo, non è stata implementata in quanto è una correttezza di tipo semantico.

Il campo **frame** può assumere solamente i valori "0", "1", "2" o ".", qualsiasi altro valore è considerato un errore. Se feature è uno dei seguenti valori: "CDS","start_codon","stop_codon" il campo frame può assumere solo i valori "0", "1" o "2", qualsiasi altro valore è considerato errore. Gli altri valori di feature disponibili devono assumere nel campo frame il valore ".", qualsiasi altro valore è considerato errore.

Tutti i campi **attributes** di ogni record hanno gli stessi due attributi obbligatori alla fine del record:

- gene_id
- transcript_id

Gli attributi devono terminare con un punto e virgola che deve quindi essere separato dall'inizio di ogni attributo successivo esattamente da uno spazio.
Gli attributi testuali sono racchiusi tra virgolette doppie.
La mancanza di questi due attributi viene segnalata come errore.

Il campo **comments** contiene i commenti che iniziano con un cancelletto "#" e continuano fino alla fine della riga. Niente oltre a "#" verrà analizzato. Questi possono verificarsi ovunque nel file, anche alla fine di una linea caratteristica separati da uno spazio o da un TAB. I commenti non sono obbligatori.
Si considera corretto anche un'intera riga di commento.

Altre spiegazioni aggiuntive sono fornite nel Jupyter Notebook.

Per eseguire i test è necessario decommentare il file gtf relativo, commentando l'attuale.

File test disponibili:

- test1.gtf contiene l'errore nel numero dei campi
- test2.gft contiene l'errore nel campo seqname
- test3.gtf contiene l'errore nel campo source
- test4.gtf contiene l'errore nel campo feature
- test5.gft contiene l'errore nei campi start ed end
- test6.gtf contiene l'errore nel campo score
- test7.gtf contiene l'errore nel campo strand
- test8.gft contiene l'errore nel campo frame
- test9.gtf contiene l'errore nel campo attributes
- test10.gtf file corretto con presenza di commenti
- test11.gtf contiene tutti i possibili errori dei file precedenti