.GTF Gene Annotations

The GTF (Gene Transfer Format) has been developed to facilitate the exchange of genome annotations (i.e., transcripts aligned to the genome) in human readable flat files. The format describes 8 (tab-separated) madatory fields, followed by an arbitrary number of key-value pairs

```
key "value"; key2 "value2"; ...
```

Mandatory Fields

Nr	Name	Value	Description
1	seqname	String	The name of the sequence. Commonly, this is the chromosome ID or contig ID. Note that the coordinates used must be unique within each sequence name in all GTFs for an annotation set.
2	source	String	Not used in the Flux Simulator: the source column should be a unique label indicating where the annotations came from --- typically the name of either a prediction program or a public database.
3	feature	String	The Flux Simulator only uses the feature "exon" which generically describes any transcribed exon.
4	start	Integer	Integer start and end coordinates of the feature relative to the beginning of the sequence named in *seqname*. *start* must be less than or equal to *end*. Sequence numbering starts at 1. Values of *start* and *end* that extend outside the reference sequence are technically acceptable, but they are discouraged.
5	end	Integer	
6	score	Integer	Not used in the Flux Simulator: the score field indicates a degree of confidence in the feature's existence and coordinates.
7	strand	[+-]	Strand of the feature, for exons/introns/transcripts corresponding their transcription directionality, either '+' or '-'.
8	frame	[012]	Not used in the Flux Simulator: 0 indicates that the feature begins with a whole codon at the 5' most base. 1 means that there is one extra base (the third base of a codon) before the first whole codon and 2 means that there are two extra bases (the second and third bases of the codon) before the first codon. Note that for reverse strand features, the 5' most base is the *end* coordinate.

Optional fields

The Flux Simulator requires the key "transcript_id" to identify exons of the same transcripts. As in the UCSC standard, transcript IDs have to be unique within the chromosome a certain transcript has been annotated on.
The automated sorting of gtf files requires the transcript_id to be in the same column across all the gtf file. This column is guessed by the first lines of the file and later on assumed to be consistent. If the column "transcript_id" varies within the gtf file, the automated sorting will fail. Such files may be fixed by the command

```
awk 'BEGIN{FS="\t";OFS="\t"}{split($NF,a," ");pfx="";s="";for(i=1;i<=length(a);i+=2){if(a[i]=="transcript_id")
{pfx=a[i]" "a[i+1]}else{s=s" "a[i]" "a[i+1]}}if(pfx==""){print "[WARN] line "NR" without transcript_id!" > "/dev
/stderr"}else{$NF=pfx""s;print$0} }' genes.gtf > genes_clean.gtf
```

or similar, where "genes.gtf" is the file with incosistent transcript_id columns and "genes_clean.gtf" is the file after reordering to match the transcript_id information in a consistent position.