.GTF format

The AStalavista parser requires WUSTL specification of the GTF v2.2 format.

- **Purpose**

GTF (Gene Transfer Format) has been designed to interchange exon-intron structures of genes. It extends the earlier defined GFF (General Feature Format) by additional fields.

- **Structure**

GTF is a tab-separated format, with each line describing a respective feature (e.g., exon, intron, CDS, start/stop codon, ..) using the following fields:

```
<seqname> <source> <feature> <start> <end> <score> <strand>  <frame> [attributes] [comments]
```

Each `attribute` is a pair of: `identifier "value";`
Textual attributes should be surrounded by doublequotes. Attributes must end in a semicolon which must then be separated from the start of any subsequent attribute by exactly one space character. Commonly used identifiers are for instance `gene_id`, `transcript_id` and `exon_id`.

Optional `comments` are ignored by AStalavista.

- **Field description**

<seqname>

The FPC (fingerprint contig) ID from the Golden Path. The field is mandatory and needed for the correct clustering of transcripts in AStalavista.

<source>

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. The field is ignored by AStalavista.

<feature>

GTF defines the following features: "`CDS`", "`start_codon`", "`stop_codon`" and "`exon`". AStalavista requires the feature "`exon`" for obtaining the exon-intron structure of each transcript. The feature "`CDS`" may be provided optionally to describe the coding sequence starting with the first translated codon and proceeding to the last translated codon. Unlike Genbank annotation, the stop codon is not included in the CDS for the terminal exon. All other features (e.g., "`start_codon`" or "`stop_codon`") are ignored by AStalavista.

<start> <end>

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 and that extend outside the reference sequence are technically acceptable, but they are discouraged for purposes of this project.

<score>

The score field (a float value) is not be used in AStalavista, so it may be replaced by a dot.

<frame>

Nucleotides to go from the `start` position of the current features to match the first position of the next codon. The feature is ignored by AStalavista.

gene_id

An unique identifier for the gene the corresponding feature is assigned to. AStalavista performs its own transcript clustering procedure and therefore ignores gene identifiers provided in the attribute list.

transcript_id

An unique identifier for the transcript the corresponding feature is assigned to. This attribute is mandatory for the correct functioning of AStalavista.

exon_id

An unique identifier for the exon the corresponding feature is assigned to. AStalavista generates exons in a cluster of transcripts non-redundantly, i.e., exons with identical `start/stop` coordinates are regarded as identical even if their `exon_ids` and/or `transcript_ids` differ.

- **Example**

Here is an example in which the "`exon`" and the "`CDS`" feature are used. The GTF lines describe a 5 exon transcript with 3 translated exons.

```
AB000381 gene_id  exon      150   200   .  +   .  gene_id "AB000381.000"; transcript_id "AB000381.000.1";
AB000381 gene_id  exon      300   401   .  +   .  gene_id "AB000381.000"; transcript_id "AB000381.000.1";
AB000381 gene_id  CDS       380   401   .  +   0  gene_id "AB000381.000"; transcript_id "AB000381.000.1";
AB000381 gene_id  exon      501   650   .  +   .  gene_id "AB000381.000"; transcript_id "AB000381.000.1";
AB000381 gene_id  CDS       501   650   .  +   2  gene_id "AB000381.000"; transcript_id "AB000381.000.1";
AB000381 gene_id  exon      700   800   .  +   .  gene_id "AB000381.000"; transcript_id "AB000381.000.1";
AB000381 gene_id  CDS       700   707   .  +   2  gene_id "AB000381.000"; transcript_id "AB000381.000.1";
```