gffpandas: a Python library to filter gff3 file entries

true   true

XX June 2018

Summary

gffpandas is a python library, which facilitates the work with general feature format version 3 (gff3) files in regard to filter the desired entries of the file. It does so by reading in the gff3 file into a data frame, based on the Python library pandas.

A gff3 file contains information about the location and attributes of genes and transcript features. It is always written in the same format, which has a header with meta-information and nine columns with the feature information (Stein 2013). With the gffpandas library it is possible to return desired entries of a gff3 file, as for example all entries of a specific feature type or within a specific base pair (bp)-length. It is also possible to split the attribute column into separate columns, splited by the attribute-tags. The data frame can then be saved again as gff3 file or even as csv or tsv file. All methods of gffpandas can be combined.

Further options of the gffpandas library are described in the project documentation (K. Förstner 2018).

The library was implemented on Ubuntu 17.10 with Python 3.6 and is dependent on the Python library pandas.

References

K. Förstner, V. Monzon. 2018. https://github.com/konrad/gffpandas.

Stein, L. 2013. https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md.