Metagenomic framework Documentation

Release 0.3.2

Francesco Rubino

May 22, 2018

Contents

Contents:

CHAPTER 1

Introduction

The aim of this library[8] is to provide a series of useful modules and packages to make it easier to build custom pipelines for metagenomics or any kind of bioinformatics analysis. It integrates other well known python libraries in bioinformatics, like HTSeq[1], pysam[2], numpy[3] and scipy[4].

A tutorial pipeline is provided in the Documentation[5].

A discussion mailing list is available at mgkit-users[6].

1.1 Citing

Rubino, F. and Creevey, C.J. 2014. MGkit: Metagenomic Framework For The Study Of Microbial Communities. . Available at: figshare[7] [doi:10.6084/m9.figshare.1269288].‘

a citation is also available using the `mgkit.cite()` function or using the *–cite* option on all scripts included. For example:

```
$ blast2gff --cite

 _|      _|    _|_|_|  _|    _|  _|    _|
 _|_|  _|_|  _|        _|  _|        _|_|_|_|
 _|  _|  _|  _|  _|_|  _|_|      _|    _|
 _|      _|  _|    _|  _|  _|    _|    _|
 _|      _|    _|_|_|  _|    _|  _|      _|_|

MGKit Version: 0.3.2

Rubino, F. and Creevey, C.J. (2014).
MGkit: Metagenomic Framework For The Study Of Microbial Communities.
```

(continues on next page)

[8] https://bitbucket.org/setsuna80/mgkit
[1] http://www-huber.embl.de/users/anders/HTSeq/
[2] https://code.google.com/p/pysam/
[3] http://www.numpy.org
[4] http://www.scipy.org
[5] http://pythonhosted.org//mgkit/pipeline/tutorial.html
[6] https://groups.google.com/forum/#!forum/mgkit-users
[7] http://figshare.com/articles/MGkit_Metagenomic_Framework_For_The_Study_Of_Microbial_Communities/1269288

(continued from previous page)

```
Available at: http://figshare.com/articles/MGkit_Metagenomic_Framework_For_The_
↪Study_Of_Microbial_Communities/1269288

[doi:10.6084/m9.figshare.1269288]
```

1.2 Links

CHAPTER 2

Installation

2.1 Docker Instance (with Jupyter Notebook)

A preconfigured Docker instance (user: mgkit, no password) has been configured using the instructions in *Installing on Ubuntu Server 16.04*, including more packages for testing, available at Docker Hub (frubino/mgkit):

```
$ docker run -p 8881:8888 -v host-dir:/home/mgkit/notebooks/ -it frubino/mgkit
```

This command (assuming that Docker is already installed), will pull the instance and present with a bash terminal. IPython and Jupyter are installed and can be used. For a recap of the options:

- *-p 8888:8888* instruct to open the port 8888 on the host
- *-v* mount the directory *host-dir* into the virtual machine */home/mgkit/notebooks/* directory
- *-it* opens an interactive shell

The port to open is port 8881 on the host for using the Jupyter Notebook. The port can be change to what best fits the user. The same applies to the working directory.

After entering the virtual machine prompt, the command that is to be used to launch the notebook is:

```
$ jupyter notebook --ip=0.0.0.0
```

After which in a browser it will possible to access it at *localhost:8881*.

Note: The Dockerfile used to build the instance is available in the repository at: docs/source/extra/Dockerfile

2.2 Requirements

The library is written completely in Python[9] and has been tested with both version 2.6.x and 2.7.x. It has not been tested with Python 3.x, but the authors are trying to write code that is potentially runnable with the use of the *2to3* tool.

[9] http://www.python.org

Most UNIX systems provide a version of Python installed. Latest versions of MacOSX provides Python 2.7.x, while most Linux variants may have different versions installed, sometimes version 3.x, but they usually provide python 2.7.x packages (Archlinux[10] uses version 3.x by default, but a *python2* package is provided). You can check the Python version installed with:

```
$ python --version
```

The library requires these Python packages:

The installation dependencies are flexible, with only *numpy* (version 1.9.2 or later) as being **required**:

```
$ pip install mgkit
```

To install every needed packages, you can use:

```
$ pip install mgkit[full]
```

Other options can be found in the *setup.py* file, in the *extra_requires* dictionary.

The optional dependencies includes:

- scipy
- pandas
- matplotlib
- pysam
- argparse (if Python 2.6 is installed, part of the standard library from 2.7)
- joblib: for script *translate_seq*, to use multiple processors
- HTSeq
- semidbm
- pymongo
- goatools[11] (required by `mgkit.mappings.go`), has package *fisher* as a dependency
- rpy2 >= 2.3.8 (required by `mgkit.utils.r_func`)

2.3 Installing on Ubuntu Server 16.04

You'll need to install the following packages with *apt-get*:

```
$ apt-get install -y velvet bowtie2 python-pip python \
  virtualenv python-dev zlib1g-dev libblas-dev \
  liblapack-dev gfortran libfreetype6-dev libpng-dev \
  fontconfig pkg-config
```

Create a virtual environment to ensure that the correct library versions are installed as explained in *Using virtualenv*.

2.4 Using pip

All dependencies are usually installed either through a package system provided by the running OS or through the pip[12] installer. If you're using a system that's shared with other people, you may not be able to install the

[10] https://www.archlinux.org/
[11] https://github.com/tanghaibao/goatools
[12] http://www.pip-installer.org/

dependencies system-wide, in which case the *–user* option in *pip* may solve the problem[24].

A system-wide installation with *pip* can be done with:

```
$ pip install path/to/library
```

while a user install is done with:

```
$ pip install --user path/to/library
```

all requirements will be downloaded/installed.

2.4.1 Using virtualenv

virtualenv[13] is a system that is used to isolate a Python installation, to make sure no conflicts arise with multiple packages. It's handy if you're developing or testing an application/library, as it provides a clean environment.

Assuming you've already installed *virtualenv*, a virtual environment can be created with:

```
$ virtualenv -p python2 mgkit-env
```

which creates a virtual environment in *mgkit-env*, with the interpreter used being the one linked to *python2*. Activating the environment requires using:

```
$ source mgkit-env/bin/activate
```

assuming you're in the same directory where you created the environment. The pip packager is installed by default with it, so we're going to use it to install the library if you have downloaded it already:

```
$ pip install path/to/library
```

or getting the last version from PyPI[14]:

```
$ pip install mgkit
```

You can also install a specific version:

```
$ pip install mgkit==0.2.0
```

2.4.2 Using the repository

The source code can also be obtained from the Bitbucket repository[15].

2.5 Running Tests

The tests requires the *nosetests* package:

```
$ pip install nose
```

and the package *yanc* is used for coloring the output. If you don't want to install it you can edit the *setup.cfg* and *setup.py* files in the source distribution and delete the *with-yanc* before running the tests.

You can run the tests with:

[24] http://www.pip-installer.org/en/latest/user_guide.html#user-installs
[13] http://www.virtualenv.org/
[14] https://pypi.python.org/pypi
[15] https://bitbucket.org/setsuna80/mgkit

```
$ python setup.py nosetests
```

Some test won't be run if the required library/data is not found. Consult the output for more information.

2.6 Building Documentation

Needs sphinx >=1.2.2

- sphinx_rtd_theme
- actdiag
- sphinxcontrib-actdiag
- blockdiag
- sphinxcontrib-blockdiag
- sphinxcontrib-napoleon (we'll be part of sphinx 1.3, needed until then)
- sphinx-argparse

Other libraries:

- graphviz
- latex (for pdf output - *make latexpdf*)

2.7 Troubleshooting

Some of the dependencies require available compilers to finish the installation. At the mimimum a system that provides the full GNU compiler suite, including a fortran compiler is required to install those dependencies by source.

If a compilation error is raised during installation, it's adviced to install each dependency manually.

I'll try to keep this section updated, but there's not that many OS that I can keep working on (mostly MacOSX and GNU/Linux).

2.7.1 HTSeq

Sometimes HTSeq or numpy fails to install in a clean environment; it's advised to install numpy first:

```
$ pip install numpy
```

and then reissue the library installation:

```
$ pip install path/to/library
```

2.7.2 MacOSX

The version of MacOSX is 10.9 that comes with Python 2.7 installed. To install every dependency from source, however it's needed to install the *Xcode* app from the **App Store** which install the compilers, with the exception of *gfortran*. Another solution is using Homebrew[16] or Macports[17], to install the compilers needed.

[16] http://brew.sh
[17] http://www.macports.org

If you want to use Xcode, you need to install the gfortran compiler, with the package provided here[18]. This should be enough to install most packages from source.

Warning: There seems to be a problem with *pandas* version 0.13.1 on MacOSX, with a segmentation fault happening when using DataFrames. The 0.14.1 version is the one tested.

Note: if there's a problem building a python package because of a compile error, dealing with an unknown command line option, use:

```
export ARCHFLAGS=-Wno-error=unused-command-line-argument-hard-error-in-future
```

It's related to the clang toolchain included with Xcode

Scipy

There are different solutions available if you have trouble installing the dependencies on MacOSX, one of which is hosted on this page[19], but installing from source is another option, provided that the Xcode and gfortran are installed.

Matplotlib

The tricky package to install in MacOSX is actually matplotlib[20], with one of many solutions being posted on a disccusion on stackoverflow[21]. In our case, installing *freetype2* and *libpng* through Homebrew it's the less painful:

```
$ brew install libpng freetype2
```

Note: If you get a compilation error which refers to freetype2 in the */opt/X11/* I found it easy to delete XQuartz installing matplotlib and then reinstall XQuartz.

Or use:

```
export PKG_CONFIG_PATH=/usr/local/Cellar/freetype/2.6_1/lib/pkgconfig/:/usr/local/
↪Cellar/libpng/1.6.19/lib/pkgconfig/
```

Note that the versions may be different.

2.7.3 Installing Scipy from source on Linux

A full description on how to install the scipy on Linux from source can be found at this address[22], be aware that the compilation of the *math-atlas* and *lapack* libraries takes a long time.

Installation in a virtual environment:

```
# create virtual environment, if needed, otherwise activate the one desired
virtualenv venv
source venv/bin/activate
# create temporary directory to compile math-atlas and lapack
```

(continues on next page)

[18] http://gcc.gnu.org/wiki/GFortranBinariesMacOS
[19] http://fonnesbeck.github.io/ScipySuperpack/
[20] http://matplotlib.org
[21] http://stackoverflow.com/questions/4092994/unable-to-install-matplotlib-on-mac-os-x
[22] http://www.scipy.org/scipylib/building/linux.html

(continued from previous page)

```
mkdir dep-build; cd dep-build
wget http://www.netlib.org/lapack/lapack.tgz
wget http://sourceforge.net/projects/math-atlas/files/Stable/3.10.2/atlas3.10.2.
↪tar.bz2/download
tar xfvj download
cd ATLAS
mkdir build; cd build
../configure -Fa alg -fPIC --with-netlib-lapack-tarfile=../../lapack.tgz --prefix=
↪$VIRTUAL_ENV
make
cd lib; make shared; make ptshared; cd ..
make install
```

This will compile math-atlas with full lapack support in the virtual environment; change the *–prefix=$VIRTUAL_ENV* to *–prefix=$HOME* if you want to install the dependencies in you home directory.

2.8 Notes

Not all packages are required to use the part of the library, but it's recommended to install all of them. Requirements are bound to change, but pandas, scipy, numpy, pysam and matplotlib are the bases of the library.

To avoid problems with the system installation, I suggest using the excellent virtualenv[23]. This will avoid problems with installing packages system-wide and breaking a working installation.

[23] http://www.virtualenv.org/

CHAPTER 3

Metagenomic Pipeline

This section detailed information about example pipelines made using the framework

3.1 Tutorial

The aim of this tutorial is to show how to build a pipeline to analyse metagenomic samples. Moreover, the SNPs calling part was made to show how diversity estimates can be calculated from metagenomic data, hence it should be changed to be more strict.

We're going to use Peru Margin Subseafloor Biosphere[25] as an example, which can be download from the ENA website.

For a pipeline using another approach, you can refer to the *HMMER Tutorial* section of the documentation. This tutorial is expected to run on a UNIX (Linux/MacOSX/Solaris), with the bash shell running, because of some of the loops (not tested with other shells).

Note: We assume that all scripts/commands are run in the same directory.

Warning: It is advised to run the tutorial on a cluster/server: the memory requirements for the programs used are quite high (external to the library).

3.1.1 Initial setup

We will assume that the pipeline and it's relative packages are already installed on the system where the tutorial is run, either through a system-wide install or a virtual environment (advised). The details are in the *Installation* section of the documentation.

Also for the rest of the tutorial we assume that the following software are installed and accessible system-wide:

- Velvet[26]

[25] https://www.ebi.ac.uk/metagenomics/project/SRP000183
[26] https://www.ebi.ac.uk/~zerbino/velvet/

- Bowtie 2[27]
- samtools[28]
- Picard Tools[29][35]
- GATK[30][36]
- HTSeq[31]
- BLAST[32] or RAPSearch2[33]

3.1.2 Getting Sequence Data

The data is stored on the EBI ftp as well, and can be downloaded with the following command (on Linux):

```
$ wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR001/SRR001326/SRR001326.fastq.gz
$ wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR001/SRR001325/SRR001325.fastq.gz
$ wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR001/SRR001323/SRR001323.fastq.gz
$ wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR001/SRR001322/SRR001322.fastq.gz
```

on MacOSX you can replace *wget* with *curl -O*.

And then uncompress with:

```
$ gunzip *.fastq.gz
```

Taxonomy Data

We only need the taxonomy for an optional part of the gene prediction for the analysis. It can be downloaded using the command:

```
$ download_data -x -p -m EMAIL
```

Where *EMAIL* should be replaced by your email address. The data will be saved in the directory *mg_data* to which we'll refer from now on.

3.1.3 Metagenome Assembly

We're going to use velvet to assemble the metagenomics sample, using the following commands in *bash*:

```
$ velveth velvet_work 31 -fmtAuto  *.fastq
$ velvetg velvet_work -min_contig_lgth 50
```

The contigs are in the *velvet_work/contigs.fa* file. We want to take out some of the informations in each sequence header, to make it easier to identify them. We decided to keep only *NODE_#*, where # is a unique number in the file (e.g. from *>NODE_27_length_157_cov_703.121033* we keep only *>NODE_27*). We used this command in bash:

```
$ cat velvet_work/contigs.fa | sed -E 's/(>NODE_[0-9]+)_.+/\1/g' > assembly.fa
```

[27] http://bowtie-bio.sourceforge.net/bowtie2/
[28] http://samtools.sourceforge.net
[29] http://picard.sourceforge.net
[35] Picard Tools needs to be found in the directory picard-tools in the same directory as this tutorial.
[30] http://www.broadinstitute.org/gatk/
[36] GATK directory is expected to be called *GATK* and inside the tutorial directory. It also needs java v1.7.x in newer versions.
[31] http://sourceforge.net/p/htseq/code/HEAD/tree/
[32] http://www.ncbi.nlm.nih.gov/books/NBK279690/
[33] http://omics.informatics.indiana.edu/mg/RAPSearch2/

3.1.4 Gene Prediction

Gene prediction can be done with any software that supports the tab format that BLAST outputs. Besides BLAST, RAPSearch can be used as well.

Before that a suitable DB must be downloaded. In this tutorial we'll use the SwissProt portion of *Uniprot <http://www.uniprot.org>* that can be downloaded using the following commands:

```
$ wget ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/
↪complete/uniprot_sprot.fasta.gz
$ gunzip uniprot_sprot.fasta.gz
```

Using BLAST

BLAST needs the DB to be indexed using the following command:

```
$ makeblastdb -dbtype prot -in uniprot_sprot.fasta
```

After which BLAST can be run:

```
$ blastx -query assembly.fasta -db uniprot_sprot.fasta -out \
    assembly.uniprot.tab -outfmt 6
```

Using RAPSearch

RAPSearch is faster than BLAST, while giving similar results. As with BLAST, there is a command to be executed before it can predict genes:

```
$ prerapsearch -d uniprot_sprot.fasta -n uniprot_sprot
```

After this command is complete its execution, RAPSearch can be started:

```
$ rapsearch -q assembly.fasta -d uniprot_sprot -o assembly.uniprot.tab
```

RAPSearch will produce two files, *assembly.uniprot.tab.m8* and *assembly.uniprot.tab.aln*. *assembly.uniprot.tab.m8* is the file in the correct format, so we can rename it and remove the other one:

```
$ rm assembly.uniprot.tab.aln
$ mv assembly.uniprot.tab.m8 assembly.uniprot.tab
```

3.1.5 Create the GFF

After BLAST or RAPSearch are finished, we can convert all predictions to a GFF file:

```
$ blast2gff uniprot -b 40 -db UNIPROT-SP -dbq 10 assembly.uniprot.tab \
    assembly.uniprot.gff
```

And then, because the number of annotations is high, we filter them to reduce the number of overlapping annotations:

```
$ filter-gff overlap assembly.uniprot.gff assembly.uniprot-filt.gff
```

This will result in a smaller file. Both script supports piping, so they can be used together, for example to save a compressed file:

```
$ blast2gff uniprot -b 40 -db UNIPROT-SP -dbq 10 assembly.uniprot.tab | \
    filter-gff overlap | gzip > assembly.uniprot-filt.gff.gz
```

Warning: filter-gff may require a lot of memory, so it's recommended to read its documentation for strategies on lowering the memory requirements for big datasets

Taxonomic Refinement

This section is optional, as taxonomic identifiers are assigned using Uniprot, but it can result in a better identification. It requires the the *nt* database from NCBI to be found on the system, in the *ncbi-db* directory.

To do it, first the nucleotide sequences must be extracted and then use blastn against the *nt* database:

```
$ get-gff-info sequence -f assembly.fasta assembly.uniprot.gff \
    assembly.uniprot.frag.fasta
$ blastn -query assembly.uniprot.frag.fasta -db ncbi-db/nt -out \
    assembly.uniprot.frag.tab -outfmt 6
```

After BLAST completes, we need to download a supporting file to associate the results with the taxonomic information:

```
$ wget ftp://ftp.ncbi.nih.gov/pub/taxonomy/gi_taxid_nucl.dmp.gz
```

and run the script to add the taxonomic information to the GFF file, also using the LCA algorithm:

```
$ add-gff-info taxonomy -v -t gi_taxid_nucl.dmp.gz -b \
    assembly.uniprot.frag.tab -s 40 -d NCBI-NT -l -x \
    mg_data/taxonomy.pickle \
    assembly.uniprot.gff assembly.uniprot-taxa.gff
```

after it completes, it is safe to rename the output GFF:

```
$ mv assembly.uniprot-taxa.gff assembly.uniprot.gff
```

Complete GFF

To add the remaining information, mapping to KO[34] and others, including the taxonomic information, a script is provided that downloads this information into a GFF file:

```
$ add-gff-info uniprot -v --buffer 500 -t -e -ec -ko \
    assembly.uniprot.gff assembly.uniprot-final.gff
```

After which we can rename the GFF file:

```
$ mv assembly.uniprot-final.gff assembly.uniprot.gff
```

3.1.6 Alignment

The alignment of all reads to the assembly we'll be made with *bowtie2*. The first step is to build the index for the reference (out assembly) with the following command:

```
$ bowtie2-build assembly.fasta assembly.fasta
```

and subsequently start the alignment, using bowtie2 and piping the output SAM file to *samtools* to convert it into BAM files with this command:

[34] http://www.kegg.jp

```
for file in *.fastq; do
    BASENAME=`basename $file .fastq`
    bowtie2 -N 1 -x assembly.fasta -U $file \
    --rg-id $BASENAME --rg PL:454 --rg PU:454 \
    --rg SM:$BASENAME | samtools view -Sb - > $BASENAME.bam;
done
```

We'll have BAM files which we need to sort and index:

```
for file in *.bam; do
    samtools sort $file `basename $file .bam`-sort;
    rm $file;
    mv `basename $file .bam`-sort.bam $file
    samtools index $file;
done
```

3.1.7 Coverage and SNP Info

The coverage information is added to the GFF and needs to be added for later SNP analysis, including information about the expected number of synonymous and non-synonymous changes. The following lines can do it, using one of the scripts included with the library:

```
$ export SAMPLES=$(for file in *.bam; do echo -a `basename $file .bam`,$file ;done)
$ add-gff-info coverage $SAMPLES assembly.uniprot.gff | add-gff-info \
    exp_syn -r assembly.fasta > assembly.uniprot-update.gff

$ mv assembly.uniprot-update.gff assembly.uniprot.gff
$ unset SAMPLES
```

The first line prepares part of the command line for the script and stores it into an environment variable, while the last command unsets the variable, as it's not needed anymore. The second command adds mapping and taxonomy information from Uniprot IDs to Kegg Orthologs, EC numbers and eggNOG.

3.1.8 SNP Calling

Before running samtools, which we'll use to do the SNP calling and GATK, to merge the vcf files, the reference *assembly.fa* must be indexed with Picard Tools (tested on version 1.30):

```
$ java -jar picard-tools/picard.jar CreateSequenceDictionary R=assembly.fasta␣
↪O=assembly.dict
```

SAMtools

For calling SNPs, we're going to use SAMtools, as it's the one having lower requirements for this tutorial. The output required by SNPdat and the later part of this tutorial is a vcf, so any software that can output can be used.

Running samtools to make the SNP calling requires a simple loop, as follows:

```
for file in *.bam; do
    samtools mpileup -Iuf assembly.fasta \
    $file | bcftools view -vcg - > `basename $file`.vcf;
done
```

After which, you need to merge all sample vcf files into one, so it can be analysed with the sample specific information, which is needed by the library. This can be done with various packages, but here we'l use GATK (tested on version 3.0-0-g6bad1c6):

```
$ export SAMPLES=$(for file in *.bam.vcf; do echo -V:`basename $file .bam.vcf`
↪$file ;done)
$ java -Xmx10g -jar GATK/GenomeAnalysisTK.jar \
  -R assembly.fasta -T CombineVariants -o assembly.vcf \
  -genotypeMergeOptions UNIQUIFY \
  $SAMPLES
$ unset SAMPLES
```

3.1.9 Data Preparation

Diversity Analysis

To use diversity estimates (pN/pS) for the data, we need to first first is aggregate all SNP information from the vcf file into data structures that can be read and analysed by the library. This can be done using the included script *snp_parser*, with this lines of bash:

```
$ export SAMPLES=$(for file in *.bam; do echo -m `basename $file .bam`;done)
$ snp_parser -v -g assembly.uniprot.gff -p assembly.vcf -a assembly.fasta $SAMPLES
$ unset SAMPLES
```

Count Data

To evaluate the abundance of taxa and functional categories in the data we need to produce one file for each sample using htseq-count, from the HTSeq library.

```
for file in *.bam; do
    htseq-count -f bam -r pos -s no -t CDS -i uid -a 8 \
    -m intersection-nonempty $file assembly.uniprot.gff \
    > `basename $file .bam`-counts.txt
done
```

Additional Downloads

The following files needs to be downloaded to analyse the functional categories in the following script:

```
$ wget http://eggnog.embl.de/version_3.0/data/downloads/COG.members.txt.gz
$ wget http://eggnog.embl.de/version_3.0/data/downloads/NOG.members.txt.gz
$ wget http://eggnog.embl.de/version_3.0/data/downloads/COG.funccat.txt.gz
$ wget http://eggnog.embl.de/version_3.0/data/downloads/NOG.funccat.txt.gz
```

and this for Enzyme Classification:

```
$ wget ftp://ftp.expasy.org/databases/enzyme/enzclass.txt
```

3.1.10 IPython Notebook

The IPython notebook with the data analysis is in the *Explore Data*. A converted python script is included in *Explore Data Python Script*

3.1.11 Full Bash Script

```
#!/bin/bash

#download data
#50 meters
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR001/SRR001326/SRR001326.fastq.gz
#1 meter
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR001/SRR001325/SRR001325.fastq.gz
#32 meters
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR001/SRR001323/SRR001323.fastq.gz
#16 meters
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR001/SRR001322/SRR001322.fastq.gz
#uncompress data
gunzip -v *.fastq.gz

#assembly - preparatory phase
velveth velvet_work 31 -fmtAuto *.fastq
#assembly
velvetg velvet_work -min_contig_lgth 50
#change sequence names
cat velvet_work/contigs.fa | sed -E 's/(>NODE_[0-9]+)_.+/\1/g' > assembly.fasta
#remove velvet working directory
rm -R velvet_work

#To use the LCA option and other analysis we need a taxonomy file
download_data -x -p -m YOUR@EMAIL

#Uniprot SwissProt DB
wget ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/
↪complete/uniprot_sprot.fasta.gz
#Uncompress it
gunzip uniprot_sprot.fasta.gz

########
#Gene prediction

###BLAST
#index Uniprot
#makeblastdb -dbtype prot -in uniprot_sprot.fasta
#Run blastx
#blastx -query assembly.fasta -db uniprot_sprot.fasta -out \
#        assembly.uniprot.tab -outfmt 6

###RAPSearch
#Index
prerapsearch -d uniprot_sprot.fasta -n uniprot_sprot
#Run
rapsearch -q assembly.fasta -d uniprot_sprot -o assembly.uniprot.tab
#rename .m8 file to assembly.uniprot.tab and delete .aln
rm assembly.uniprot.tab.aln
mv assembly.uniprot.tab.m8 assembly.uniprot.tab

########
#Converts gene prediction into GFF annotations
blast2gff uniprot -b 40 -db UNIPROT-SP -dbq 10 assembly.uniprot.tab \
        assembly.uniprot.gff
filter-gff overlap assembly.uniprot.gff assembly.uniprot-filt.gff
#rename the new filtered file
mv assembly.uniprot-filt.gff assembly.uniprot.gff

########
#Taxonomic refinement - requires NCBI nt DB installed and indexed
export NCBINT_DIR=ncbi-db
```

(continues on next page)

(continued from previous page)

```
if [ -d "$NCBINT_DIR" ]; then
        echo "Taxonomic refinement";
        #Extract annotations sequences
        get-gff-info sequence -f assembly.fasta assembly.uniprot.gff \
                assembly.uniprot.frag.fasta
        #Use blastn to match against NCBI NT
        blastn -query assembly.uniprot.frag.fasta -db ncbi-db/nt -out \
                assembly.uniprot.frag.tab -outfmt 6

        #Download necessary data
        wget ftp://ftp.ncbi.nih.gov/pub/taxonomy/gi_taxid_nucl.dmp.gz

        add-gff-info taxonomy -v -t gi_taxid_nucl.dmp.gz -b \
                assembly.uniprot.frag.tab -s 40 -d NCBI-NT -l -x \
                mg_data/taxonomy.pickle \
                assembly.uniprot.gff assembly.uniprot-taxa.gff
        #rename the file to continue the script
        mv assembly.uniprot-taxa.gff assembly.uniprot.gff
fi
unset NCBINT_DIR

########
#Finalise information from Gene Prediction
#Adds remaining taxonomy, EC numbers, KO and eggNOG mappings
add-gff-info uniprot -v --buffer 500 -t -e -ec -ko \
        assembly.uniprot.gff assembly.uniprot-final.gff
#Rename the GFF
mv assembly.uniprot-final.gff assembly.uniprot.gff

########
#Alignments
bowtie2-build assembly.fasta assembly.fasta
for file in *.fastq; do
        BASENAME=`basename $file .fastq`
        bowtie2 -N 1 -x assembly.fasta -U $file \
        --very-sensitive-local \
        --rg-id $BASENAME --rg PL:454 --rg PU:454 \
        --rg SM:$BASENAME | samtools view -Sb - > $BASENAME.bam;
done
#sort and index BAM files with samtools
for file in *.bam; do
        samtools sort $file `basename $file .bam`-sort;
        #removes the unsorted file, it's not needed
        rm $file;
        mv `basename $file .bam`-sort.bam $file
        samtools index $file;
done

########
#Add coverage and expected changes to GFF file
export SAMPLES=$(for file in *.bam; do echo -a `basename $file .bam`,$file ;done)
#Coverage info
add-gff-info coverage $SAMPLES assembly.uniprot.gff | add-gff-info \
        exp_syn -r assembly.fasta > assembly.uniprot-update.gff
#rename to continue the script
mv assembly.uniprot-update.gff assembly.uniprot.gff
unset SAMPLES

########
#SNP calling using samtools
for file in *.bam; do
```

(continues on next page)

(continued from previous page)

```
        samtools mpileup -Iuf assembly.fasta \
        $file | bcftools view -vcg - > `basename $file`.vcf;
done

#Index fasta with Picard tools - GATK requires it
java -jar picard-tools/picard.jar CreateSequenceDictionary \
        R=assembly.fasta O=assembly.dict

#merge vcf
export SAMPLES=$(for file in *.bam.vcf; do echo -V:`basename $file .bam.vcf` $file;done)
java -Xmx10g -jar GATK/GenomeAnalysisTK.jar \
          -R assembly.fasta -T CombineVariants -o assembly.vcf \
          -genotypeMergeOptions UNIQUIFY \
          $SAMPLES
unset SAMPLES

########
#snp_parser
export SAMPLES=$(for file in *.bam; do echo -m `basename $file .bam`;done)
snp_parser -v -g assembly.uniprot.gff \
        -p assembly.vcf \
        -a assembly.fasta \
         $SAMPLES
unset SAMPLES

########
#Count reads
for file in *.bam; do
        htseq-count -f bam -r pos -s no -t CDS -i uid -a 8 \
        -m intersection-nonempty $file assembly.uniprot.gff \
        > `basename $file .bam`-counts.txt
done

########
#eggNOG mappings
wget http://eggnog.embl.de/version_3.0/data/downloads/COG.members.txt.gz
wget http://eggnog.embl.de/version_3.0/data/downloads/NOG.members.txt.gz
wget http://eggnog.embl.de/version_3.0/data/downloads/COG.funccat.txt.gz
wget http://eggnog.embl.de/version_3.0/data/downloads/NOG.funccat.txt.gz

########
#EC names
wget ftp://ftp.expasy.org/databases/enzyme/enzclass.txt
```

3.1.12 Explore Data Python Script

3.2 HMMER Tutorial

This example pipeline explore three different aspects from the *Tutorial*):

1. normalisation of metagenomic data using khmer[37]
2. the use of another assembler, MEGAHIT[38]
3. Using Kegg to identify the ortholog genes from the nitrogen metabolism
4. making custom HMMER profiles

[37] http://khmer.readthedocs.org/
[38] https://github.com/voutcn/megahit

5. the use of samtools/bcftools for SNP calling

Hint: The normalisation assembly and profile building steps take a long time, with high relatively high memory requirements. Moreover, the profile building requires an active network connection. The complete assembly is available in this tutorial data, as well as the built HMM profile. These can be used if you are stuck on one of those steps.

3.2.1 Requirements

Warning: The requirements for assembly/normalisation are high for a desktop computer, with the khmer normalisation step using ~6GB of RAM to complete with a reasonable low false positive rate. The computer tested was running Mac OS X v10.11, with 16GB of RAM.

Table 1: Software Tested

Software	Version
wget	any
velvet	1.2.10
bowtie2	2.2.6/2.1.0
khmer	2.0
hmmer	3.1b2/3.1b1
clustalo	1.2.1
megahit	1.0.3
samtools	1.2.0
bcftools	1.0

Table 2: Python Packages

Package	Version
mgkit	0.2.1
HTSeq	0.6.1
pandas	0.17.1
pysam	0.8.4
scipy	0.16.1
semidbm	0.5.1
matplotlib	1.5
seaborn	0.6

On Mac OS X, some of the software requirements can be installed using Homebrew, with this command:

```
$ brew install wget homebrew/science/velvet homebrew/science/bowtie2 \
homebrew/science/samtools pyenv-virtualenv homebrew/science/hmmer \
homebrew/science/clustal-omega
```

Note: a lot of the shell code is possible in Bash, so you need to make sure the correct shell is loaded.

3.2.2 Metagenomic Data

The data that will be used is from the Analysis of metagenome in a full scale tannery wastewater treatment[39] project on ENA[40]. It has 5 samples, from waste water management:

- I (Influent)
- B (Buffering)
- SA (Secondary aeration)
- PA (Primary aeration)
- SD (Sludge digestion)

As it involves waste water management, it's interesting to understand the diversity of the genes involved in the nitrogen metabolism.

The raw reads files can be downloaded with the following command:

```
$ wget ftp://ftp.sra.ebi.ac.uk/vol1/ERA315/ERA315794/fastq/B_R1.fastq.gz \
ftp://ftp.sra.ebi.ac.uk/vol1/ERA315/ERA315794/fastq/B_R2.fastq.gz \
ftp://ftp.sra.ebi.ac.uk/vol1/ERA315/ERA315794/fastq/I_R1.fastq.gz \
ftp://ftp.sra.ebi.ac.uk/vol1/ERA315/ERA315794/fastq/I_R2.fastq.gz \
ftp://ftp.sra.ebi.ac.uk/vol1/ERA315/ERA315794/fastq/PA_R1.fastq.gz \
ftp://ftp.sra.ebi.ac.uk/vol1/ERA315/ERA315794/fastq/PA_R2.fastq.gz \
ftp://ftp.sra.ebi.ac.uk/vol1/ERA315/ERA315794/fastq/SA_R1.fastq.gz \
ftp://ftp.sra.ebi.ac.uk/vol1/ERA315/ERA315794/fastq/SA_R2.fastq.gz \
ftp://ftp.sra.ebi.ac.uk/vol1/ERA315/ERA315794/fastq/SD_R1.fastq.gz \
ftp://ftp.sra.ebi.ac.uk/vol1/ERA315/ERA315794/fastq/SD_R2.fastq.gz
```

3.2.3 Digital Normalisation and QC

One aspect that makes the assembly of metagenomic particularly complex is the different coverage of the different organisms that are found in a sample. One solution is the use of a kmer counting such as **khmer**, as it allows to reduce the differences. This also reduces the memory requirements of the assembly.

The first step is to interleave the paired end reads into one file, which can be done using the *interleave-reads.py* script from the *khmer* package. As the quality, observer using FastQC[41] is poor in the last 20 bp, the following bash code uses Python and HTSeq to cut the last 20 bp from both reads files, using IO streams. This avoids the use of temporary files, but it's not mandatory.

```
$ interleave-reads.py --gzip -o all-interleaved.fq.gz \
<(
python - <<END
import HTSeq, sys, glob
files = glob.glob('*R1.fastq.gz')
for fname in files:
    for record in HTSeq.FastqReader(fname):
        record = record[:-20] # cut 20 bp
        # HTSeq adds '[part]' to the header, the next line cut it
        record.name = record.name[:-6]
        record.write_to_fastq_file(sys.stdout)
END
) \
<(
python - <<END
import HTSeq, sys, glob
files = glob.glob('*R2.fastq.gz')
```

(continues on next page)

[39] http://www.ebi.ac.uk/ena/data/view/PRJEB6461
[40] http://www.ebi.ac.uk/ena/
[41] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

(continued from previous page)

```
for fname in files:
    for record in HTSeq.FastqReader(fname):
        record = record[:-20]
        record.name = record.name[:-6]
        record.write_to_fastq_file(sys.stdout)
END
)
```

Note: The interleave passage is especially important since some khmer scripts had problems parsing the new Casava header of Fastq files.

The file *all-interleaved.fq.gz* will be created, containing the trimmed sequences, interleaved.

The digital normalisation can be done using the *normalize-by-median.py*, part of the *khmer* package. In this dataset, with the parameters used, around 7% of data can be excluded from the assembly:

```
$ normalize-by-median.py -k 24 -p -o normalised.fq.gz \
  --gzip -M 6e9 all-interleaved.fq.gz
```

Producing a *normalised.fq.gz* file that can be used with an assembler.

3.2.4 Assembly

The assembly, as usual can be done with any assembler, as long as its output is a FASTA file. MEGAHIT can be used to assembler this data using the following command:

```
$ megahit --presets meta --verbose --min-contig-len 100 \
  --12 normalised.fq.gz -o megahit-out
```

One things that can create problems in software such as samtools, is the presence of spaces in the sequence headers. The following python code (executed in a BASH shell) can be used to give unique names to each sequence and keep in a file the names assigned by the assembler for any future reference:

```
$ python - <<END
from mgkit.io import fasta
from uuid import uuid4
import json

seq_dict = {}
with open('final-contigs.fa', 'w') as f:
    for name, seq in fasta.load_fasta('megahit-out/final.contigs.fa'):
        uid = str(uuid4())
        seq_dict[uid] = name
        fasta.write_fasta_sequence(f, uid, seq)

json.dump(seq_dict, open('seq-dict.json', 'w'))
END
```

This small script will read all sequences from the output of MEGAHIT, the *final.contigs.fa* file and replaces the original sequence headers with new ones, using the *uuid4* function. This function generate random unique identifiers without spaces. A dictionary with the original sequence headers is saved as a JSON file.

3.2.5 Download Data

It is necessary to download the taxonomy to download the genes from Kegg (for this tutorial). Moreover the taxonomy will be used in later steps and to download it, the MGKit script *download_data* can be used:

```
$ download_data -p -x -m EMAIL
```

This will save a *taxonomy.pickle* file in the subdirectory *mg_data*. More information at the script documentation (*download-data - Download Taxonomy from NCBI*)

3.2.6 Create Profiles

To download only a portion of Kegg, in this case the genes from the nitrogen metabolism, it's needed to use the Kegg REST API to retrieve the gene IDs. MGKit include a client class for this in the `mgkit.kegg` module, called *KeggClientRest*. Its method *link_ids* returns a data structure that contains the list of genes in the Nitrogen Metabolism (ID: ko00910).

The script can be executed on a bash shell with the following command:: $ export GENES='python - <<END from mgkit import kegg k = kegg.KeggClientRest() print " ".join(k.link_ids('ko', 'ko00910')['ko00910']) END'

The command will export an environment variable called *GENES* that can be passed to the *download_profiles* script:

```
$ download_profiles -o profiles -ko $GENES -r order -l bacteria \
-m EMAIL -t mg_data/taxonomy.pickle
```

The script will retrieve all the sequences in the *profiles* directory, for the genes in the *GENES* environment variable. The genes will be download as one for file for each order of bacteria AND gene. Each file will contain, all sequences for that ortholog, found in Uniprot for a particular order of bacteria. Orders with just one sequence will be skipped.

After the script finish its execution, the *GENES* variable can be unset:

```
$ unset GENES
```

More information about the script is in *download-profiles - Download Custom Profiles*.

Alignment and HMM profiles

The files created must be aligned using Clustal Omega or another similar software and for each alignment a HMM profile can be built using the *hmmbuild* from the HMMER package. An example, using a BASH loop:

```
$ for file in profiles/*.fa; do
        echo $file `basename $file .fa`
    clustalo -i $file -o profiles/`basename $file .fa`.afa ;
    hmmbuild profiles/`basename $file .afa`.hmm profiles/`basename $file .fa`.afa;
done
```

Each single profile can then concatenated using *cat*:

```
$ cat profiles/*.hmm > profiles.hmm
```

3.2.7 Gene Prediction

With the HMM profiles in one file, *hmmsearch* can be used to searc for similarity in the assembly. Before that, *hmmsearch* can only work on aminoacid sequences, so the assembly must be translated in the possible frames. The recommended way to do this is to use the *translate_seq* script included with MGKit:

```
$ translate_seq final-contigs.fa final-contigs.aa.fa
```

Warning: The problem with other software to translate the assembly is that the script that convert the result of *hmmsearch* into a GFF file needs the information about the frame and strand. *translate_seq* append a suffix to the sequence header to indicate it. As an example, for a sequence named *contig0001*, the following sequences headers will be found: *contig0001-f0*, *contig0001-f1*, *contig0001-f2*, *contig0001-r0*, *contig0001-r1*, *contig0001-r2*. The *f* stands for the forward (+ strand), *r* for the reverse (- strand), the number indicates the frame, from 1 to 2.

hmmsearch can then be launched using the following command:

```
$ hmmsearch -o /dev/null --domtbl hmmer_dom-table.txt profiles.hmm final-contigs.
↪aa.fa
```

The only file needed by MGKit is the domain table, stored in the file *hmmer_dom-table.txt*.

GFF Creation

The output of *hmmsearch* can then be supplied to the *hmmer2gff* script included in MGKit, which converts it to a GFF file:

```
$ hmmer2gff -d -o assembly.gff final-contigs.aa.fa hmmer_dom-table.txt
```

The command will create a *assembly.gff* file from all hits in the domain table. In this case the e-value filter was disabled (*-d* option), because the collection of files may be too small.

3.2.8 GFF Filtering

The GFF filtering works in the same way as explained in *Tutorial* and more informations can be found in the script manual (*filter-gff - Filter GFF annotations*). One thing to point out is that most scripts and commands in MGKit (and other software) allow the use of pipes, concatenating multiple commands in one line to avoid the use of temporary files. The following command can be run on a BAST shell:

```
$ cat assembly.gff | filter-gff values -b 40 | filter-gff \
overlap -s 100 | \ add-gff-info kegg -c EMAIL \
-v -d > assembly.filt.gff
```

The commands do the following, in sequence (between |):

1. output to the standard output the content of *assembly.gff*
2. only keeps annotations that have a bit score of at least 40
3. filters overlapping annotations (for at least 100 bp), keeping the one with the highest bit score
4. add the names of the genes getting them from Kegg

The result is then outputted into the *assembly.filt.gff* (after the >). This is not enforced, but can be used to speed up some commands, as nothing is written to disk, and avoid confusion when managing multiple temporary files.

3.2.9 GFF Additions

At this point, we have most of the information we need to continue with the analysis, but there is one mandatory and one optional step which can be done. The GFF after filtering can is not enough to continue with the diversity analysis, as we need coverage information about each predicted gene. We can also refine the taxonomic assignment, which is detailed in *Tutorial*.

Computing the gene coverage can be done before filtering, but the number of annotations would be too high, so it's preferred to add coverage information after filtering the GFF. Also, because we needs alignment files for each sample to compute the gene coverage, it is advised to makes the alignment files in parallel with the GFF filtering, to speed up the pipeline.

Alignments

To create alignments for each sample, the assembly file must first be indexed, with the folowing command:

```
$ bowtie2-build final-contigs.fa final-contigs
```

The following BASH loop creates the BAM file for each sample:

```
$ for file in *R1.fastq.gz; do
        BASENAME=`basename $file _R1.fastq.gz`
        echo $file $BASENAME "$BASENAME"_R2.fastq.gz
        bowtie2 -N 1 -x final-contigs --local --sensitive-local \
        -1 $file -2 "$BASENAME"_R2.fastq.gz \
        --rg-id $BASENAME --rg PL:Illumina --rg PU:Illumina-MiSeq \
        --rg SM:$BASENAME --no-unal \
        2> $BASENAME.log | samtools view -Sb - > $BASENAME.bam;
done
```

The loop uses the list of reads file from the first element of the pairs (R1.fastq.gz files) to automate the process, resulting in files that are in the form *SAMPLE.bam* (e.g. B.bam). A log file is also kept, using the same file name and *log* extension.

The alignments made need to be sorted, and the following BASH loop give an example of this:

```
$ for file in *.bam; do
        samtools sort -T tmp.$file -O bam -o `basename $file .bam`-sort.bam $file;
        mv `basename $file .bam`-sort.bam $file;
        samtools index $file;
done
```

Note: samtools 1.2 (at least) needs to specify the format and temp file prefix. Later version may not require it.

The result is BAM files with the sample names, as before.

Coverage and Expected SNPs

The alignments can now be used also to add coverage information to the GFF file, which is needed for another script in the pipeline. Because sample names are needed, as the per sample coverage information is store as *sample_cov*, adding a suffix *_cov* after the sample name, the following command infers the sample names from the BAM files:

```
$ export SAMPLES=$(for file in *.bam; do echo -a `basename $file .bam`,$file ;done)
```

The *SAMPLES* environment variable is used for the script *add-gff-info*, in particular its *coverage* command:

```
$ add-gff-info coverage $SAMPLES assembly.filt.gff | \
add-gff-info exp_syn -r final-contigs.fa > assembly.filt.cov.gff
```

To also add the information about the expected number of synonymous and non-synonymous changes for each annotation, the *exp_syn* command for the *add-gff-info* was used. The file is then saved as *assembly.filt.cov.gff*.

The *SAMPLES* variable can now be unset:

```
$ unset SAMPLES
```

3.2.10 SNP Calling

In this tutorial, it was decided to use samtools/bcftools to call SNPs, as GATK can require too much memory and time. The following command calls SNPs for all samples, writing a *assembly.vcf* file to disk:

```
$ samtools mpileup -t DP,SP,DPR,DV -ugf final-contigs.fa *.bam \
| bcftools call -vmO v > assembly.vcf
```

Note: samtools version 1.2 and bcftools version 1.0 were tested

Using the VCF file created, the *snp_parser* script included in MGKit can be used:

```
export SAMPLES=$(for file in *.bam; do echo -m `basename $file .bam`;done)

snp_parser -s -v -g assembly.filt.cov.gff -p assembly.vcf \
-a seqs/final-contigs.fa $SAMPLES

unset SAMPLES
```

The *SAMPLES* variable is dynamically created to help write the command line for *snp_parser* and after its execution can be unset. The command line is different compared to *Tutorial*, as *-s* was added to distinguish the type of sample information in the VCF, as outputted by *bcftools*, compared to one created using GATK (the default type for *snp_parser*).

3.2.11 IPython Notebook

The IPython notebook with the data analysis is in the *Explore Data*. A converted python script is included in *Explore Data Python Script*

3.2.12 Full Bash Script

```
#!/bin/bash

# Some tools require a contact email
export EMAIL=your@email

# Requirements
#
# software          - version tested
# mgkit          - 0.2.1
# wget                  - any
# velvet          - 1.2.10
# bowtie2          - 2.2.6 / 2.1.0
# khmer          - 2.0
# hmmer          - 3.1b2 / 3.1b1
# clustalo          - 1.2.1
# megahit         - 1.0.3

# python packages
# HTSeq>=0.6.1
# pandas>=0.17.1
# pysam>=0.8.4
# scipy>=0.16.1
# semidbm>=0.5.1
# matplotlib>=1.5
# seaborn>=0.6

# Requirements specific for Mac OS X (El Capitan, 10.11), uncomment these lines
# brew install wget homebrew/science/velvet homebrew/science/bowtie2 \
# homebrew/science/samtools pyenv-virtualenv homebrew/science/hmmer \
# homebrew/science/clustal-omega

```

(continues on next page)

(continued from previous page)

```
# Memory ~6 GB for khmer normalisation
# ~3.5 GB for megahit
# ~0.5 GB for bowtie2

# Using data from the following project:
# http://www.ebi.ac.uk/ena/data/view/PRJEB6461
# Samples:
# I (Influent)
# B (Buffering)
# SA (Secondary aeration)
# PA (Primary aeration)
# SD (Sludge digestion)
mkdir seqs
cd seqs
#download data
wget ftp://ftp.sra.ebi.ac.uk/vol1/ERA315/ERA315794/fastq/B_R1.fastq.gz ftp://ftp.
↪sra.ebi.ac.uk/vol1/ERA315/ERA315794/fastq/B_R2.fastq.gz ftp://ftp.sra.ebi.ac.uk/
↪vol1/ERA315/ERA315794/fastq/I_R1.fastq.gz ftp://ftp.sra.ebi.ac.uk/vol1/ERA315/
↪ERA315794/fastq/I_R2.fastq.gz ftp://ftp.sra.ebi.ac.uk/vol1/ERA315/ERA315794/
↪fastq/PA_R1.fastq.gz ftp://ftp.sra.ebi.ac.uk/vol1/ERA315/ERA315794/fastq/PA_R2.
↪fastq.gz ftp://ftp.sra.ebi.ac.uk/vol1/ERA315/ERA315794/fastq/SA_R1.fastq.gz ftp:/
↪/ftp.sra.ebi.ac.uk/vol1/ERA315/ERA315794/fastq/SA_R2.fastq.gz ftp://ftp.sra.ebi.
↪ac.uk/vol1/ERA315/ERA315794/fastq/SD_R1.fastq.gz ftp://ftp.sra.ebi.ac.uk/vol1/
↪ERA315/ERA315794/fastq/SD_R2.fastq.gz

# Normalisation of the reads
# Khmer when using paired reads works best when are interleaved
# Moreover, it requires the older Casava type of header
# The sequences are trimmed by 20 bp and fed to the khmer
# script interleave-reads.py (the quality is not good in the last 20 bp)
interleave-reads.py --gzip -o all-interleaved.fq.gz \
<(
python - <<END
import HTSeq, sys, glob
files = glob.glob('*R1.fastq.gz')
for fname in files:
    for record in HTSeq.FastqReader(fname):
        record = record[:-20] # cut 20 bp
        # HTSeq adds '[part]' to the header, the next line cut it
        record.name = record.name[:-6]
        record.write_to_fastq_file(sys.stdout)
END
) \
<(
python - <<END
import HTSeq, sys, glob
files = glob.glob('*R2.fastq.gz')
for fname in files:
    for record in HTSeq.FastqReader(fname):
        record = record[:-20]
        record.name = record.name[:-6]
        record.write_to_fastq_file(sys.stdout)
END
)
# Normalisation takes out almost 7% of the reads
normalize-by-median.py -k 24 -p -o normalised.fq.gz --gzip -M 6e9 all-interleaved.
↪fq.gz
# rm all-interleaved.fq.gz

# megahit manages to assemble the data
# add -t N to use more processors
```

(continues on next page)

(continued from previous page)

```
megahit --presets meta --verbose --min-contig-len 100 --12 normalised.fq.gz -o megahit-out
# megahit used spaces in the sequence headers so it may create problems later
# one solution is to assign to each contig a new random sequence and then
# keep track of those in a json dictionary for later (if needed)
python - <<END
from mgkit.io import fasta
from uuid import uuid4
import json

seq_dict = {}
with open('final-contigs.fa', 'w') as f:
    for name, seq in fasta.load_fasta('megahit-out/final.contigs.fa'):
        uid = str(uuid4())
        seq_dict[uid] = name
        fasta.write_fasta_sequence(f, uid, seq)

json.dump(seq_dict, open('seq-dict.json', 'w'))
END
rm -R megahit-out
cd ..

#download offline data (only taxonomy)
download_data -p -x -m $EMAIL

# Get the nitrogen metabolism KOs
export GENES=`python - <<END
from mgkit import kegg
k = kegg.KeggClientRest()
print " ".join(k.link_ids('ko', 'ko00910')['ko00910'])
END`

download_profiles -o profiles -ko $GENES -r order -l bacteria -m $EMAIL -t mg_data/taxonomy.pickle

unset GENES

#Profile alignments and build
for file in profiles/*.fa; do
        echo $file `basename $file .fa`
    clustalo -i $file -o profiles/`basename $file .fa`.afa ;
    hmmbuild profiles/`basename $file .afa`.hmm profiles/`basename $file .fa`.afa;
done
#Make one file for all profiles
cat profiles/*.hmm > profiles.hmm

#translation of the assembly in AA
translate_seq seqs/final-contigs.fa final-contigs.aa.fa

#HMMER command line
# add --cpu N to use more processors
hmmsearch -o /dev/null --domtbl hmmer_dom-table.txt profiles.hmm final-contigs.aa.fa
#convert into annotations
hmmer2gff -d -o assembly.gff final-contigs.aa.fa hmmer_dom-table.txt
# Filter annotations and add information from Kegg
# most scripts working on GFF files allow to pipe their input/output
# First remove annotations with a bit score less than 40, then filter
# overlapping annoations and finally add gene descriptions and pathways for
# for each annotations. `cat` is not necessary, since the input/output can
# specified with a `-` (dash), which is standard
```

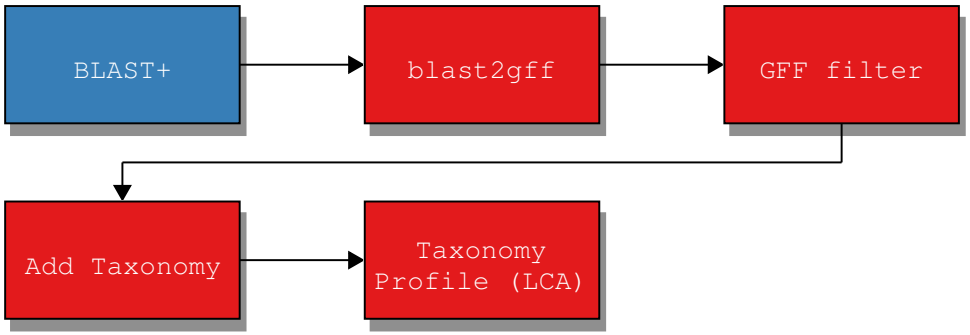
(continues on next page)

(continued from previous page)

```
cat assembly.gff | filter-gff values -b 40 | filter-gff overlap -s 100 | \
add-gff-info kegg -c $EMAIL -v -d > assembly.filt.gff

# bowtie2
# index
bowtie2-build seqs/final-contigs.fa final-contigs
# alignments, read groups are added, using a sensitive approach for alignment
# the reads that are not aligned are not included in the BAM files (--no-unal
# option)
for file in seqs/*R1.fastq.gz; do
        BASENAME=`basename $file _R1.fastq.gz`
        echo $file $BASENAME seqs/"$BASENAME"_R2.fastq.gz
        bowtie2 -N 1 -x final-contigs --local --sensitive-local \
        -1 $file -2 seqs/"$BASENAME"_R2.fastq.gz \
        --rg-id $BASENAME --rg PL:Illumina --rg PU:Illumina-MiSeq \
        --rg SM:$BASENAME --no-unal \
        2> $BASENAME.log | samtools view -Sb - > $BASENAME.bam;
done

# samtools
# sorting (by position) the BAM files and indexing them
for file in *.bam; do
        # samtools 1.2 (at least) needs to specify the format and temp file prefix
        samtools sort -T tmp.$file -O bam -o `basename $file .bam`-sort.bam $file;
        mv `basename $file .bam`-sort.bam $file;
        samtools index $file;
done

#index for the assembly (required by GATK and samtools)
#.fai index
samtools faidx seqs/final-contigs.fa

#add coverage data
########
#Add coverage and expected changes to GFF file
export SAMPLES=$(for file in *.bam; do echo -a `basename $file .bam`,$file ;done)
#Coverage info
add-gff-info coverage $SAMPLES assembly.filt.gff | add-gff-info \
        exp_syn -r seqs/final-contigs.fa > assembly.filt.cov.gff
unset SAMPLES

# samtools
# -u uncompressed
# -g binary BCF (it's piped to bcftools)
# -f reference fasta
# -t several important per sample information
# bcftools
# call - new command (instead of view)
# -v - vcf output
# -m - type of calling
# -O v - uncompressed vcf
samtools mpileup -t DP,SP,DPR,DV -ugf seqs/final-contigs.fa *.bam \
| bcftools call -vmO v > assembly.vcf

#snp_parser
export SAMPLES=$(for file in *.bam; do echo -m `basename $file .bam`;done)
snp_parser -v -g assembly.filt.cov.gff -p assembly.vcf \
-a seqs/final-contigs.fa -s $SAMPLES
unset SAMPLES
```

3.2.13 Explore Data Python Script

3.3 Profile a Community with BLAST



The above diagram shows the process of getting a community profile from a BLAST run against a DB of choice. The choice of DB is up to the user, but any DB that provides a NCBI *taxon_id* can be used. Such DBs include the ones provided by NCBI (e.g. nt, nr, viral) as well as Uniprot (SwissProt, TrEMBL).

The community profile will use an assembly and we want to assign each of its contigs to taxon. This can be done a BLAST output and a series of scripts that ends with the *lca* command of the *taxon_utils* script (*taxon-utils - Taxonomy Utilities*). *lca* stands for *last common ancestor*, which indicates that given a number of taxa, we try to resolve the taxon they all have in common. This can be of any level, from a specific strain to a kingdom, such as Bacteria.

There a cases when there's no *lca* that can be resolved and this is due to the way NCBI taxonomy is structured, with multiple top levels, such as *cellular organisms*, *viruses* and so on.

Note: Other DB may provide the *taxon_id* from NCBI, but this should be checked by the user

---

3.3.1 Considerations

Since the assembly of a metagenome is a time consuming process, the assembly from the *HMMER Tutorial* tutorial will be used. We'll try to use all results from the *nt* DB from NCBI, as well as separate *viruses* and *cellular organisms* and resolve the *lca* for those annotations separately.

Another thing to consider is how to filter the annotations. That's up to the user to decide which suits the specific task, but these options will be examined:

1. filter based on a static threshold, such as > 50 bitscore (using *filter-gff values*)
2. filter based on a dynamically chosen value, keeping only the X top options (using *filter-gff sequence*)
3. filter based on overlap (using *filter-gff overlap*)

The results may differ, and they are listed from the fastest to the slowest.

3.3.2 Requirements

MGKit should be installed and its scripts can be run from the command line. Refer to the installation for this, but it's assumed that it was installed with:

```
$ pip install mgkit[full]
```

Moreover, the tutorial makes use of UNIX command line utilities, so a version of GNU/Linux, *BSD or MacOS X should be used to run this tutorial. *BASH* is expected to be the shell running.

The following should be installed as well:

- BLAST+ (blastn will be used)
- ncftp (used to download the *NCBI nt* DB)

MacOS X

The software requirements can be installed with *homebrew*, using the following command:

```
$ brew install ncftp blast
```

3.3.3 Download Data

Assembly

TODO

NCBI nt

We'll be using the NCBI nt which will be stored in the *ncbi-nt* directory. If a copy is already somewhere, just create a symbolic link to that directory, for example:

```
$ ln -s path-to-the-db ncbi-nt
```

Otherwise, a copy can be downloaded and prepared with the following commands:

```
$ mkdir ncbi-nt
cd ncbi-nt
ncftpget ftp://ftp.ncbi.nlm.nih.gov/blast/db/nt*.gz
for x in *.tar.gz; do tar xfvz $x ;done
rm *.tar.gz
cd ..
```

ID to Taxonomy

The following file contains the *taxon_id* for all the IDs in the *NCBI nt* DB. It will be used to add taxonomic information before running the *lca* step:

```
$ wget ftp://ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid/nucl_gb.accession2taxid.
↪gz
```

NCBI Taxonomy

This can be installed using the following script included in MGKit:

```
$ download-taxonomy.sh
```

Which will create a file called *taxonomy.pickle*

3.3.4 Community Profiling

To make the tutorial faster, we'll filter the assembly file to include only contigs of at least 500bp:

```
$ python - <<END
from mgkit.io import fasta
with open('final-contigs-filt.fa', 'w') as f:
    for name, seq in fasta.load_fasta('final-contigs.fa'):
        if len(seq) >= 500:
            fasta.write_fasta_sequence(f, name, seq)
END
```

BLAST

The *blastn* command will be used to search for similar sequences in the *NCBI nt* DB. The following command will create a tab separated file with the results:

```
$ blastn -query final-contigs-filt.fa -db ncbi-nt/nt -outfmt 6 -out assembly-nt.
↪tab -evalue 0.001
```

Convert into a GFF

The following command will create a GFF file from the BLAST output:

```
$ blast2gff blastdb -i 3 -r assembly-nt.tab assembly-nt.gff
```

We're using the *blastdb* command of the *blast2gff* command, since it gives more control over the way the header file is formatted:

```
gi|118501159|gb|CP000482.1|
```

At the moment, the header format of the *NCBI nt* DB is a | (pipe) list that contains two type of identifiers. The first element is *gi*, to indicate that the following element (second) is the GI identifier that it's being retired in September 2016. The third is indicates the DB from where the other ID originates from (GenBank in this case) and the fourth is the identifier that we'll use.

By default *blast2gff blastdb* used the second element (*118501159*) of the header as *gene_id*, so we use:

1. *-i 3* to instead use the fourth element (*CP000482.1*) as *gene_id*
2. *-r* will remove the versioning information from the *gene_id*, so *CP000482.1* will become *CP000482*

The reason for this is that the file containing the *taxon_id* for each identifier is better used with a fourth element of the header without the versioning information.

Adding the Taxonomic Information

The *add-gff-info addtaxa* command allows to insert taxonomic information (in the GFF *taxon_id* attribute) into the GFF file. This step integrates the content of the *nucl_gb.accession2taxid.gz* file with the GFF file. The structure of this file is:

ACCESSION ACCESSION.VERSION TAXONID GI

Warning: this command has to load all the GFF in memory, so a high memory machine should be used (~30GB). The GFF can be split into smaller files to save memory and the subsection here will describe the process.

Since we used the *ACCESSION* as *gene_id*, we need to edit the file to pass it to the *add-gff-info addtaxa* command *-t* option. This can be don on the fly and the following command adds information to the GFF file created:

```
$ add-gff-info addtaxa -t <(gunzip -c nucl_gb.accession2taxid.gz | cut -f 1,3) -e␣
↪assembly-nt.gff assembly-nt-taxa.gff; mv assembly-nt-taxa.gff assembly-nt.gff
```

The *-t* option is the file that can contains the *taxon_id* for each *gene_id*, the script accept a tab separated file. After the this we rename the output file to keep less files around. The *-e* option was used to remove from the output file any annotation for which a *taxon_id* was not found. Since we need them for the LCA later, it makes sense to remove them before filtering.

Reduce Memory Usage

First we need to split the *assembly-nt.gff* file, with a good option being using the *split* command in Unix. The following command will create the files:

```
$ split -l 1000000 -d assembly-nt.gff split-gff
```

This command will create 12 GFF files (of at most 1 milion lines each), whose names start with *split-gff*. Since we split the files we can use a loop to add the taxonomic information to all of them:

```
$ for x in split-gff*; do
add-gff-info addtaxa -t <(gunzip -c nucl_gb.accession2taxid.gz | cut -f 1,3) -e $x
↪$x-taxa;
done
```

This reduces the memory usage to ~2.5GB, but it takes longer to re-read the *nucl_gb.accession2taxid.gz* 12 times. There are way to parallelise it, but they are beyond the scope of this tutorial.

After the command has finished running, the content of the files can be concatenated into a single file again and delete the split files:

```
$ cat split-gff*-taxa > assembly-nt.gff; rm split-gff*
```

Filter the GFF

As mentioned we'll provide three different way to filter a GFF, before passing it to the script that will output the *lca* information. This way we can compare the different filtering strategies.

Filter by Value

Let's assume a scenario where we're working on reads or very short contigs. We may decide to use a threshold, so the filtering is fast, but doesn't compromise the quality of the assignment. This can achieved using the *filter-gff values* command:

```
$ filter-gff values -b 50 assembly-nt.gff assembly-nt_filt-value.gff
```

The command will read the GFF file and keep only the hits that are greater than or equal to 50, which we're assuming is a good compromise for the assignment. This filtering strategy has the advantage of operating on a per-annotation basis, so the memory usage is low and no grouping or calculation is required.

Filter Dynamically

While the above can be give good results, we can think of cases where the number of hits that pass that threshold may be high (e.g. a conserved sequence in multiple organisms). In this case a more sensible choice would be to keep only the hits that are in the top 5-10% of the hits on that contig, all those over the median, mean or any other

threshold based on the distribution of a sequence's hits. The *filter-gff sequence* command can be used to filter the GFF:

```
$ filter-gff sequence -t -q .95 -c ge assembly-nt.gff assembly-nt_filt-sequence.gff
```

The options used will keep only the hits that have a bitscore (evalue and identity can also be used) greater than or equal to the top 5% of the bitscore distribution for that contig.

This threshold will include also contigs that have only one hit (that's the reason to use *-c ge* instead of *-c gt*). We also assume that the input GFF is sorted (*-t* option) by contig name, to use less memory.

Filter Ovelaps

Let's assume that in some cases the we think there may be cases where the contig contains regions that different rates of conservation. The first filter may keep too many taxa with similar sequences in a portion of the contig, while the second one may not provide enough coverage of the contig, keeping only the very best hits.

In this case, we can use the *filter-gff overlap* command to keep of all overlapping hits only the best one. And since we want to make sure that we still have good homology, we could still filter by value the hits, before that filter.

The following command will make that type of filtering:

```
$ filter-gff values -b 50 assembly-nt.gff | sort -s -k 1,1 -k 7,7 | filter-gff␣
↪overlap -t -s 1 - assembly-nt_filt-overlap.gff
```

We just chained the filtering from the *values* command, keeping only annotations with at least 50 bitscore and passing it to the sort command. This passage is not necessary it the the *-t* option is not used with *filter-gff overlap*, but it uses less memory by pre-sorting the GFF by contig/strand first, since the *filter-gff overlap* works on each strand separately. We also used the *-s* options to trigger the filter for annotations that overlap for as much as 1 bp.

More information about this type of filter can be found in *Tutorial* and *filter-gff - Filter GFF annotations*.

Getting the Profile

We'll have 3 GFF files ending in *final.gff*, one per each type of filtering, that contain the *taxon_id* for each annotation they contain.

Note: these files are available at this page[42] if you want to skip

---

Since the filtered files are available now, we can create a file that contains the LCA assignments. We can ouput 2 type of files (see *taxon-utils - Taxonomy Utilities*), but for the purpose of this tutorial, we'll get a GFF file that we can also use in a assembly viewer. The command to create them is:

```
1 $ for x in *filt-*.gff; do
2 taxon_utils lca -v -t taxonomy.pickle -r final-contigs-filt.fa -s -n `basename $x .
  ↪gff`-nolca.tab -ft LCA-`echo $x | egrep -o 'value|overlap|sequence' | tr␣
  ↪[:lower:] [:upper:]` $x `basename $x .gff`-lca.gff;
3 done
```

The options used are:

- *-t* to direct the script to the taxonomy that we already downloaded
- *-r* to output a GFF with one annotation per contig that covers the whole sequence
- *-s* indicates that the input is sorted by reference sequence
- *-n* outputs a tab separated file with the contigs that could not be assigned

[42] http://bitbucket.org

- *-ft* is used to change the *feature type* column in the GFF, from the dafault *LCA* to one which includes the type of filtering used

The file ending in *-nolca.tab* contain the contigs that could not be assigned, while the files ending in *-lca.gff* contain the taxonomic assignments, with the *taxon_id* pointing to the assigned taxon identifier, *taxon_name* for the taxon scientific name (or common name if none is found) and *lineage* contains the whole lineage of the taxon.

Using Krona

Besides having a file with the assignments and a GFF that can be used in Tablet, a quick profile can be produced using Krona[43] and its associated *Krona Tools*. To produce a file that can be used with Krona Tools the **-k** can be used with the *taxon_utils lca* command. An additional option is to give the tool the total number of sequences in the assembly with the **-kt** option, to have a complete profile of the assembly:

```
$ for x in *filt-{overlap,sequence,value}.gff; do
taxon_utils lca -v -t taxonomy.pickle -k -kt `grep -c '>' final-contigs-filt.fa` -
↪s $x `basename $x .gff`-lca.krona;
done
```

To the *-kt* option was passed the total number of sequences (just used grep to count how many headers are in the fasta file).

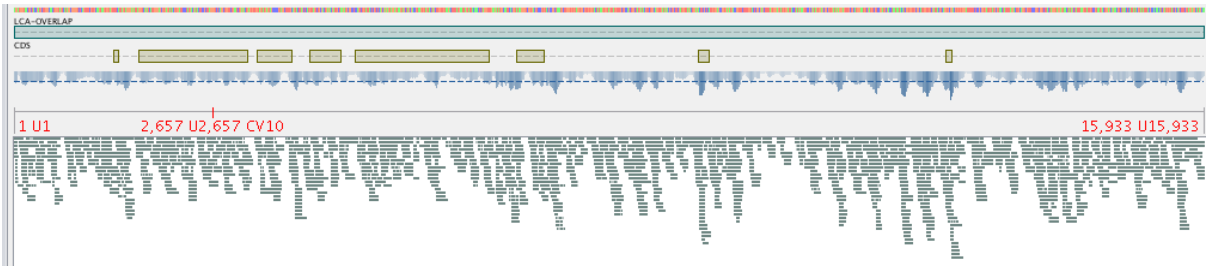
The produced files with **krona** extension can be the be used with the **ktImportText** (or **ImportText** if Krona Tools were not installed). The **-q** option of the script must be used:

```
$ for x in *.krona; do
ktImportText -q -o `basename $x .krona`.html $x;
done
```

This will create an HTML file for each one that can be read in a web browser.

Using Tablet

The GFF created can be used in software such as Tablet[44]. The image below shows a contig with the features loaded from the filtered (overlap) GFF and the GFF LCA file produced by *taxon_utils lca*.



3.4 Gene Prediction

Gene prediction is an essential portion of a metagenomic pipeline, because there is no a priori knowledge of what genes are in the samples. Moreover, a gene must be taxonomically annotated to correlate its function to the taxonomic group it belongs to.

There different ways to predict genes, with some relying on general function domains like the ones from PFam[45] or others. This type of collections is very useful in identifying proteins in an unknown sequence. The main

[43] https://github.com/marbl/Krona/wiki
[44] https://ics.hutton.ac.uk/tablet/
[45] http://pfam.xfam.org/

drawback for the examined datasets is that it's not possible to identify the organism but only the general function of a sequence.

A second approach is to use orthologs, genes derived from the same ancestral sequence with their separation originated from a speciation process. As functionality is preserved among them, they are a good choice when approaching samples where multiple organisms are present. Two collections, eggNOG[46] and Kegg Orthologs[47], are highly curated. A single ortholog gene identifier maps to several genes from different organisms, so the characterisation of an ortholog gene propagate to all its associated genes. This includes links to pathways in the case of Kegg and to functional categories in the case of eggNOG.

These genes are shared between organisms, so a single ortholog gene corresponds to several genes in different organisms. In some cases this is a preferred approach, as it allows a good resolution in the function, especially because this collections are linked to pathways in the case of Kegg and to functional categories in the case of eggNOG. The downside is that the collection of gene is not extensive and it is not connected to a taxonomic identification.

Another approach is to use genes from general public databases, like Uniprot[48]. While more general a collection, compared to Kegg Orthologs or eggNOG, it offers mappings to these two collections, as well as others. It does contain when available taxonomic information of its genes and it is divided into a manually curated portion (SwissProt) and an automated one (TrEMBL). This separation allows to have mixing annotations from both portions while preferentially use the ones from SwissProt.

In general the framework does not enforce one collection over another and in fact ortholog genes were used in one study, while Uniprot genes were used in others.

3.4.1 Prediction Software

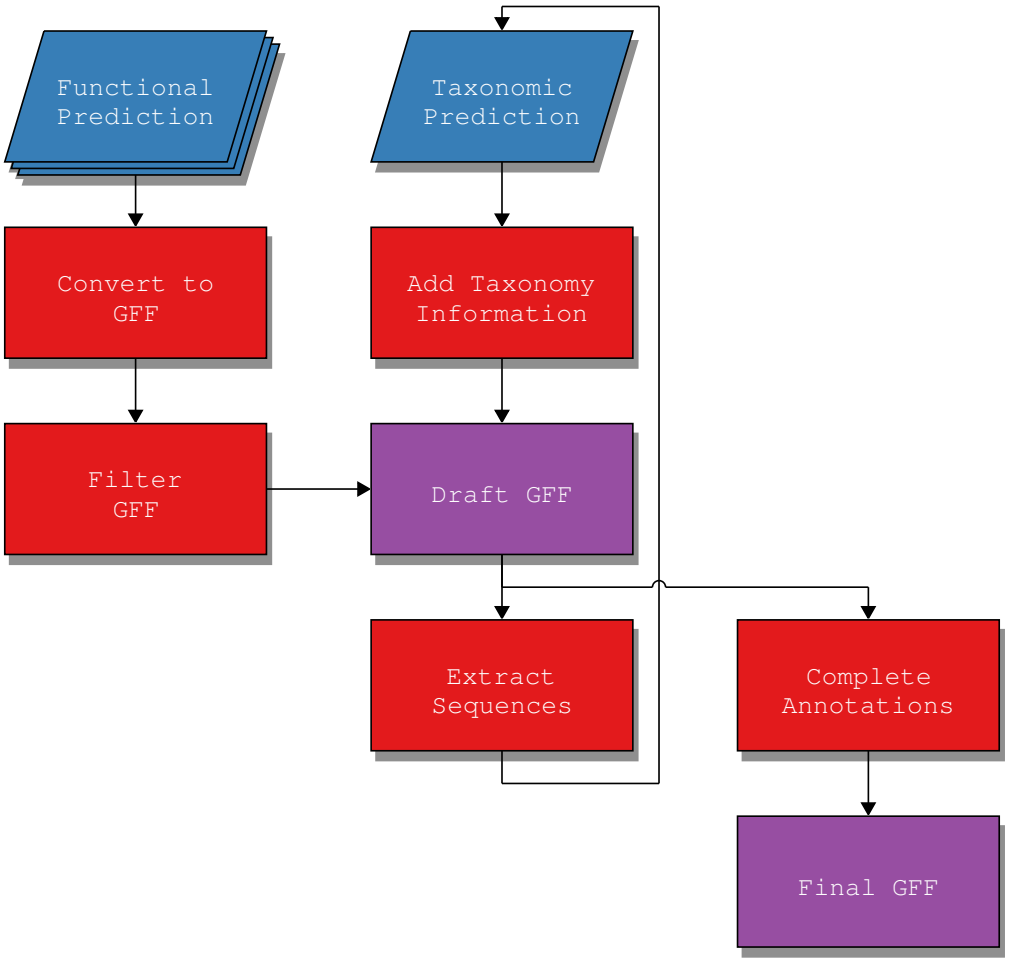
The prediction of genes requires both a collection and specific softwares to find homologous sequneces. There are two classes of software that can be used for gene prediction: one is profile based and the second uses similarity search.

An example of software using profile search is HMMER[49]. This approach uses an alignment of similar sequences to create an hidden Markov model (HMM) profile that is used to identify sequences that are similar to said profile. Curated profiles can be created from the eggNOG collection or other collection, but is also possible to automate the process of creating custom profiles.

The similarity search approach, is used in BLAST+[50], where a collection of sequences is first indexed and then all words in the index are searched against the query sequence and the most similar ones are investigated further to report a region of similarity.

[46] http://eggnog.embl.de
[47] http://www.kegg.jp/kegg/ko.html
[48] http://www.uniprot.org
[49] http://hmmer.janelia.org/
[50] http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download

3.4.2 General Procedure



The end result of the full process is a GFF file[51] including *gene_id taxon_id* and *uid* attributes. This attributes are need to identify univocally the annotation (*uid*), the gene that was functionally predicted (*gene_id*) and the organism it belongs to (*taxon_id*). A more detailed explanation of these attributes and others is in *MGKit GFF Specifications*.

The choice of the format is based on the fact that it is to manipulate without ad-hoc tools, as it is a text file, and it is accepted as input file by several bioinformatics tools.

3.4.3 Functional Prediction

The first step of the pipeline is to generate functional prediction information of the metagenomic sequences. This can be achieved using any tool of choice, with HMMER and BLAST+ preferred among others. While BLAST+ is being extensively tested, any program that outputs prediction data in the same tab separated format as BLAST+ can be used, including USEARCH[52] and RAPSearch2[53].

The framework provides two script, one for HMMER output *hmmer2gff - Convert HMMER output to GFF* and one for BLAST+ tab separated format *blast2gff - Convert BLAST output to GFF*, to convert predictions to GFF annotations.

[51] http://www.sequenceontology.org/gff3.shtml
[52] http://www.drive5.com/usearch/
[53] http://omics.informatics.indiana.edu/mg/RAPSearch2/

Usually a filter on the quality of the prediction is used, with 40 bit as a minimum indication of homology and 60 being a better one. If more than one gene collection is used, for example both SwissProt and TrEMBL, it is advised to keep track of which collection the annotation comes from and giving the chosen collection a quality index (*dbq* attribute in the GFF), with higher values assigned to preferred collections.

Generate Profiles

Note: More detailed information is in *download-profiles - Download Custom Profiles*

The framework provides, via a script and the following guidelines, a way to generate profiles for Kegg Orthologs genes, using Uniprot as repository of sequnece data.

The process of building the profiles to be used with HMMER is a step that involves several tasks:

1. download of data
2. alignment of sequences
3. conversion in HMMER profiles.

The first step involves, for all ortholog genes, to download all sequences available for each taxon level of interest: this will produce a series of file which contain the amino-acid sequences for each tuple gene-taxon. The sequences downloaded are aligned using Clustal Omega[54] and for each alignment a profile is built.

Building profiles in this way, by going through all ortholog genes and choosing the taxon level desired, opens the possibility of incrementally refining the profiling of a metagenome without having to rerun all profiles again, as only the new ones need to be run. Filtering the all the results is a much faster operation.

3.4.4 Filter Annotations

The number of predictions generated by the chosen prediction software can be very high, with a lot of them having just a few base pairs difference. When this involves the same functional prediction, it is safe to use the one with the best score.

However, when multiple genes are predicted on roughly the same region of a sequence, the choice of the annotation to keep is more difficult. Overlapping annotations can be either a weaker prediction, or the result of a chimeric sequence, as it can happen in metagenomic assemblies.

To solve this problem a script (filter-gff) was written that filters annotations when an overlap occurs. The algorithm scans the list of all annotations in a single sequence, sorted by their bit score, trying to find annotations that overlap. The filter is triggered when two annotations overlap, for at least 100bp by default, and the annotation to keep is chosen using a function that maximise three parameters: db quality (dbq), bit score (bitscore) and annotation length, in order of priority. This greatly reduces the number of annotations remaining and keeps the best possible annotations.

The choice of the 100 bp, as default value for an overlap to trigger filtering between two annotations, is based on the comparison of 36 prokaryotic genomes retrieved from UCSC[55] gene overlaps.

Table 3: Archaeal Genomes

Crenarchaea	Euryarchaea	Thaumarchaea
Acidianus hospitalis	Archaeoglobus fulgidus	Cenarchaeum symbiosum
Desulfurococcus kamchatkensis	Haloarcula marismortui	Nitrosopumilus maritimus
Hyperthermus butylicus	Methanobrevibacter ruminantium M1	
Pyrobaculum islandicum	Methanobrevibacter smithii	
Thermoproteus tenax Kra1	Thermococcus barophilus MP	
	Thermococcus onnurineus	

[54] http://www.clustal.org/omega/
[55] https://genome.ucsc.edu

Table 4: Bacterial Genomes

Actinobacteria	Aquificae	Bacteroidetes	Proteobacteria	Spirochaetes
Acidothermus cellulolyticus 11B	Aquifex aeolicus	Bacteroides thetaiotaomicron	Blochmannia floridanus	Borrelia burgdorferi
Bifidobacterium longum	Hydrogenivirga sp. 128	Cytophaga hutchinsonii	Candidatus Carsonella ruddii	Leptospira interogans
Mycobacterium tuberculosis	Hydrogenobaculum 3684	Gramella forsetii	Photobacterium profundum	Treponema pallidum
Nocardioides JS614	Persephonella marina	Salinibacter ruber	Salmonella typhi	
Rhodococcus RHA1	Sulfurihydrogenbium YO3AOP1		Shewanella oneidensis	
Tropheryma whipplei TW08 27	Sulfurihydrogenibium yellowstonense		Vibrio parahaemolyticus	

3.4.5 Taxonomic Prediction

When using Uniprot to functionally predict genes in a sequence, the metadata available for the gene may contain taxonomic information. However, while a gene from one species may have been predicted in the data, this prediction may be incorrect. There are various reasons, closely related organisms, lack of specific genes for a class of organisms or annotations, among others.

In this cases the approach taken in the framework is to extract the predicted nucleotide sequences using the tool of choice, provieded that it names the sequences using the *uid* attribute of an annotation, or the provided script (*get-gff-info - Extract informations to GFF annotations*). The sequences included in the file can be used with a similarity search program as BLAST to find the closest related sequences.

The collection used for this is the *nt* database from NCBI and a search against it can provide a better taxonomic assignment. The default behaviour is to take the taxonomic prediction with the highest score. It is recommended to use only predictions with a bit score of 60 or higher.

An included script *add-gff-info - Add informations to GFF annotations* provides the functionality necessary to add the taxonomic assignments to the GFF file. It also includes a last common ancestor (LCA) algorithm to resolve ambiguous assignments.

Last Common Ancestor

While the default behaviour is to take a prediction with the highest score, this may not be correct if more predictions have similar score. For this reason a last common ancestor (LCA) algorithm can be enabled on the predictions that are a set number of bits from the highest one, with the default value used 10.

The algorithm works by collecting all taxonomic predictions for a sequence, that falls within the chosen threshold, and traversing the taxonomy to find the last common ancestor. If no common ancestor can be found, the taxonomic predictions are discarded.

3.4.6 Complete Annotations

When a GFF file is produced by the framework, it can be integrated with the taxonomic information from Uniprot, if that was the collection used to predict genes.

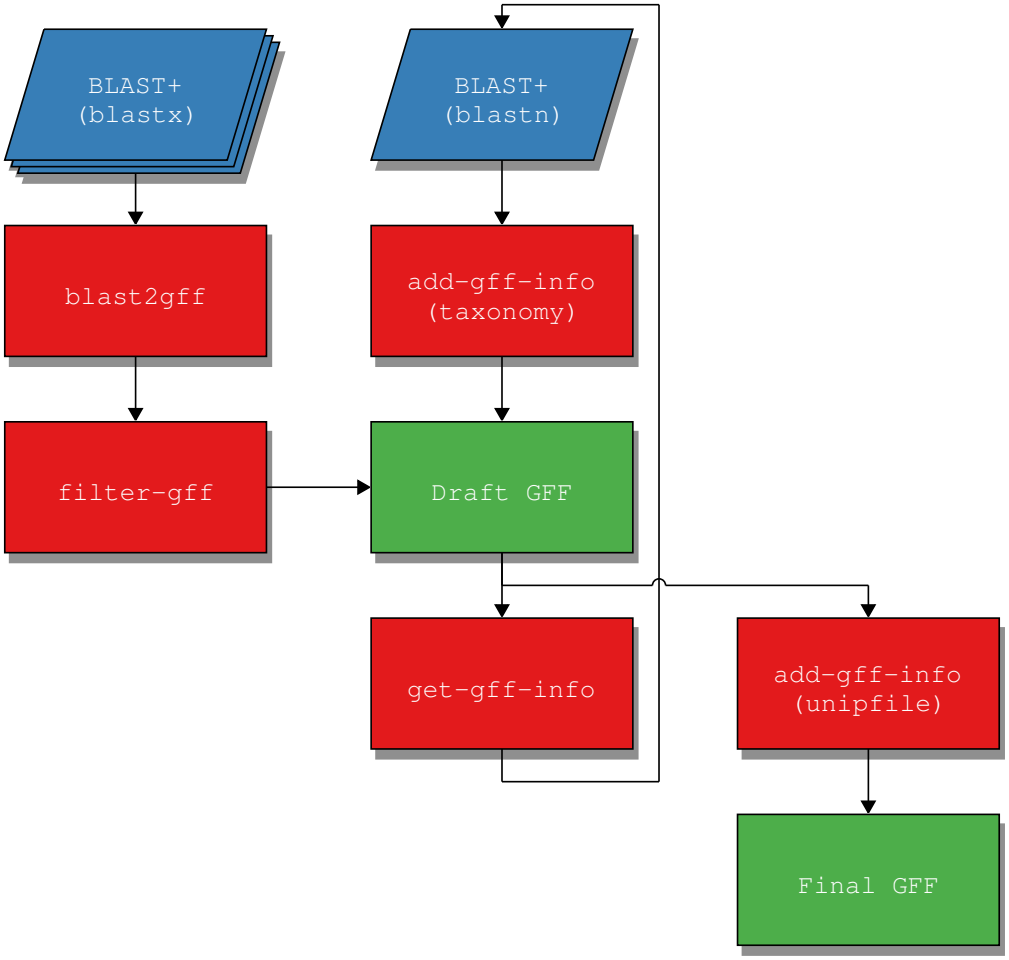
The process add mapping attributes to the GFF file, with eggNOG and Kegg Orthologs for example, while also completing the taxonomic assignment of annotations that were not assigned taxonomically. This can be done with an included script *add-gff-info - Add informations to GFF annotations* and the completed GFF can be used for further analysis

3.4.7 Examples

Gene Prediction with BLAST+

BLAST is another option to predict genes in a sequence and it is less difficult to set up as it only needs a FASTA file with the collection of genes to use.

The examples here use Uniprot DBs to predict genes, as it enables the mapping to several DBs, including eggNOG and Kegg Orthologs. It also assumes that an assembly has been produced for the gene prediction and that a DB to use blast with is already set up. Also, the BLAST+ package is expected to be installed on the system.



Functional Prediction

Assuming that BLAST is correctly installed and that the Uniprot DB is indexed, the only required parameter required by the scripts is *–outfmt 6*, which produces the BLAST tab format required by scripts that convert a BLAST output to a GFF *blast2gff - Convert BLAST output to GFF*. An example of the command line is this:

```
$ blastx -query assembly.fasta -db uniprot_sprot.fasta -out assembly.uniprot.tab -
↪outfmt 6
```

This will output a file that can be passed to the GFF creation script, *blast2gff*, with the following command:

```
$ blast2gff uniprot -b 40 -db UNIPROT-SP -dbq 10 assembly.uniprot.tab assembly.
↪uniprot.gff
```

The script documentation *blast2gff - Convert BLAST output to GFF* offer more information on the parameters. Suffice to say that *-b 40* excludes any BLAST hit with a bit score of less than 40 and *-dbq 10* point to the DB quality, as per *MGKit GFF Specifications*, which is important to filter annotations coming from multiple DBs with varying quality.

Filter GFF

The amount of prediction can be huge and most of them are overlapping annotations, so filtering the GFF annotations is important. A script is included to filter annotations (*filter-gff - Filter GFF annotations*), whose *overlap* command filters overlapping annotations. An example of the script execution is:

```
$ filter-gff overlap assembly.uniprot.gff assembly.uniprot-filt.gff
```

This will considerably reduce the size of the GFF file.

Taxonomic Prediction

Once the functional annotations are filtered, the next step is to assign taxonomic information to them, with the process being a two step process, to further refine the assignments.

The base process is to use the taxonomic assignment associated with the Uniprot ID predicted by BLAST, with a possible refinement of this by using the nucleotidic sequence associated with an annotation, whose similarity is then predicted using BLAST against a large collection of sequences, like the *nt* DB in NCBI.

Taxonomic Refinement

This part is entirely optional, but should be executed before the next one, to speed the scripts that follow.

First the sequences from the GFF file needs to be extracted with the *get-gff-info sequence* (*get-gff-info - Extract informations to GFF annotations*) command; an example execution is:

```
$ get-gff-info sequence -f assembly.fasta assembly.uniprot.gff assembly.uniprot.
↪frag.fasta
```

This will output a FASTA file called *assembly.uniprot.fasta* with the sequences used as query for the *blastn* command of the BLAST+ package against the *nt* DB:

```
$ blastn -query assembly.uniprot.frag.fasta -db nt -out assembly.uniprot.frag.tab -
↪outfmt 6
```

The ouput file *assembly.uniprot.frag.tab* is then passed to the *taxonomy* command of the *add-gff-info* script to incorporate the assignments information into the GFF file, an example of the execution of this command is the following:

```
$ add-gff-info taxonomy -t gi_taxid_nucl.dmp.gz -b assembly.uniprot.frag.tab -s 40␣
↪-d NCBI-NT assembly.uniprot.gff assembly.uniprot-taxa.gff
```

More information about the options used can be found at the script documentation (*get-gff-info - Extract informations to GFF annotations*), with an LCA option being available for assignments.

Complete Annotations

The rest of the taxonomic assignments, if not all, as well as additional informations can be added with *uniprot* or *unipfile* commands of the *add-gff-info* *add-gff-info - Add informations to GFF annotations* script. The main difference is that the *uniprot* command may be slower, as it connects to the internet and on a large number of annotations it takes a long time. The *unipfile* uses a file provided by Uniprot with additional information (in particular the taxonomy).

An example execution of the command is:

```
$ add-gff-info unipfile -i idmapping.dat.gz -m NCBI_TaxID assembly.uniprot.gff␣
↪assembly.uniprot-final.gff
```

Note: if you used the taxonomic refinement, use *assembly.uniprot-taxa.gff* instead of *assembly.uniprot.gff*

CHAPTER 4

Scripts Details

This section detailed information about the scripts included

4.1 blast2gff - Convert BLAST output to GFF

4.1.1 Overview

Blast output conversion in GFF requires a BLAST+ tabular format which can be obtained by using the *–outfmt 6* option with the default columns, as specified in `mgkit.io.blast.parse_blast_tab()`. The script can get data from the standard in and ouputs GFF lines on the standard output by default.

Uniprot

The Function `mgkit.io.blast.parse_uniprot_blast()` is used, which filters BLAST hits based on bitscore and adds by default a *db* attribute to the annotation with the value *UNIPROT-SP*, indicating that the SwissProt db is used and a *dbq* attribute with the value 10. The feature type used in the GFF is CDS.

BlastDB

If a BlastDB, such as *nt* or *nr* was used, the **blastdb** command offers some quick defaults to parse BLAST results.

It now includes options to control the way the sequence header are formatted. Options to change the separator used, as well as the column used as *gene_id*. This was added because at the moment the GI identifier (the second column in the header) is used, but it's being phased out in favour of the embl/gb/dbj (right now the fourth column in the header). This should easy the transition to the new format and makes it easier to adapt an older pipeline/blastdb to newer files (like the ID to TAXA files).

The header from the a *ncbi-nt* header looks like this:

```
gi|160361034|gb|CP000884.1
```

This is the default output accepted by the *blastdb* command. The fields are separated by | (pipe) and the GI is used (–gene-index 1, since internally the string is split by the separator and the second element is take - lists indices are 0-based in Python). This output uses the following options:

```
--header-sep '|' --gene-index 1
```

Notice the single quotes to pass the pipe symbol, since *bash* would interpret it as pipeing to the next coommand otherwise. This is the default.

In case, for the same header, we want to use the *gb* identifier, the only option to be specified is:

```
--gene-index 3
```

This will get the fourth element of the header (since we're splitting by pipe).

As in the *uniprot* command, the *gene_id* can be set to use the whole header, using the *-n* option. Useful in case the *BLAST* db that was used was custom made. While pipe is used in major databases, it was made the default, by if the db used has different conventions the separator can be changed. There's also the options of later changing the *gene_id* in the output GFF if necessary.

Changes

Changed in version 0.2.6: added *-r* option to *blastdb*

Changed in version 0.2.5: added more options to give user control to the *blastdb* command

New in version 0.2.3: added *–fasta-file* option, added more data from a blsat hit

New in version 0.2.2: added *blastdb* command

Changed in version 0.2.1: added *-ft* option

Changed in version 0.1.13.

- added *-n* parameter to *uniprot* command
- added *-k* option to *uniprot* command

New in version 0.1.12.

4.1.2 Options

Convert BLAST output to a GFF file

```
usage: blast2gff [-h] [-v | --quiet] [--cite] [--manual] [--version]
                 {uniprot,blastdb} ...
```

Named Arguments

-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

Sub-commands:

uniprot

Blast results from a Uniprot database, by default SwissProt

```
blast2gff uniprot [-h] [-db DB_USED] [-n] [-dbq DB_QUALITY] [-b BITSCORE]
                  [-k ATTR_VALUE] [-ft FEAT_TYPE] [-a FASTA_FILE]
                  [-v | --quiet] [--cite] [--manual] [--version]
                  [input_file] [output_file]
```

Positional Arguments

input_file BLAST+ output file in tabular format, defaults to stdin

Default: -

output_file Output GFF file, defaults to stdout

Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-db, --db-used Uniprot database used with BLAST

Default: "UNIPROT-SP"

-n, --no-split **if used, the script assumes that the sequence header contains** only the gene id

Default: False

-dbq, --db-quality Quality of the DB used

Default: 10

-b, --bitscore Minimum bitscore to keep the annotation

Default: 0.0

-k, --attr-value **Additional attribute and value to add to each annotation,** in the form attr:value

-ft, --feat-type Feature type to use in the GFF

Default: "CDS"

-a, --fasta-file Fasta file with nucleotide sequences, used to calculate the frame, if not used, the frame on the '-' strand will always be 0

-v, --verbose more verbose - includes debug messages

Default: 20

--quiet less verbose - only error and critical messages

--cite Show citation for the framework

--manual Show the script manual

--version show program's version number and exit

blastdb

Blast results from a NCBI database, like *nt*

```
blast2gff blastdb [-h] [-db DB_USED] [-n] [-s HEADER_SEP] [-i GENE_INDEX] [-r]
                  [-dbq DB_QUALITY] [-b BITSCORE] [-k ATTR_VALUE]
                  [-ft FEAT_TYPE] [-a FASTA_FILE] [-v | --quiet] [--cite]
                  [--manual] [--version]
                  [input_file] [output_file]
```

Positional Arguments

input_file BLAST+ output file in tabular format, defaults to stdin

Default: -

output_file Output GFF file, defaults to stdout

Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-db, --db-used blastdb used

Default: "NCBI-NT"

-n, --no-split **if used, the script assumes that the sequence header will be** used as gene_id

Default: False

-s, --header-sep The separator for the header, defaults to 'l' (pipe)

Default: "l"

-i, --gene-index **Which of the header columns (0-based) to use as gene_id** (defaults to 1 - the second column)

Default: 1

-r, --remove-version **if used, the script removes the *version* information from the** gene_id

Default: False

-dbq, --db-quality Quality of the DB used

Default: 10

-b, --bitscore Minimum bitscore to keep the annotation

Default: 0.0

-k, --attr-value **Additional attribute and value to add to each annotation,** in the form attr:value

-ft, --feat-type Feature type to use in the GFF

Default: "CDS"

-a, --fasta-file Fasta file with nucleotide sequences, used to calculate the frame, if not used, the frame on the '-' strand will always be 0

-v, --verbose more verbose - includes debug messages

Default: 20

--quiet less verbose - only error and critical messages

--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

4.2 filter-gff - Filter GFF annotations

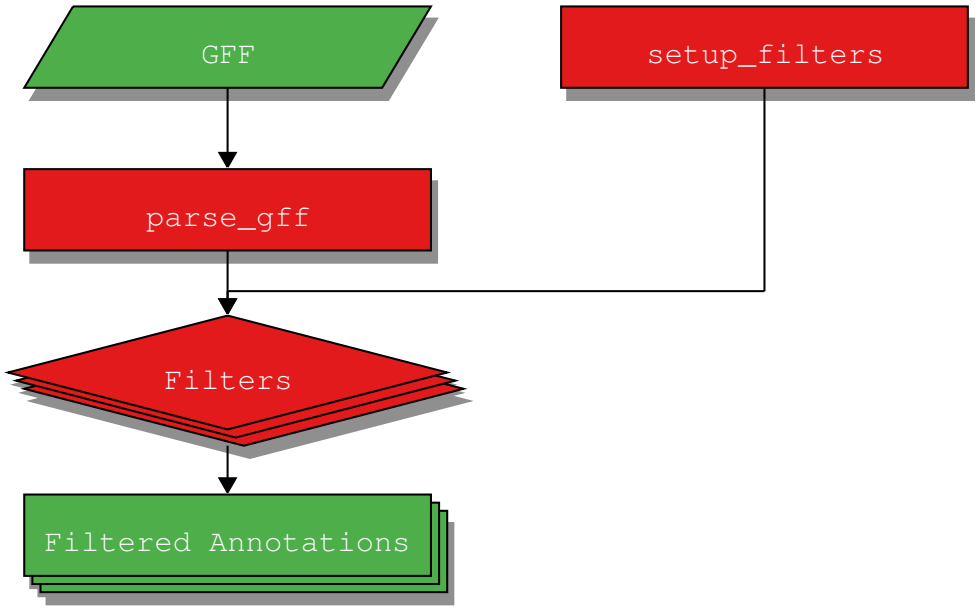
4.2.1 Overview

Filters GFF annotations in different ways.

Value Filtering

Enables filtering of GFF annotations based on the the first 8 columns, which are fixed values as well using the last column which holds information in a key=value way. There are some predefined key=value filters, like *gene_id*, but *–str-eq*, *–str-in*, *–num-ge* and *–num-le* allow to make additional filters.

The functions used to make the filters are located in the module `mgkit.filter.gff`, and their names start with *filter_base*, *filter_attr* and *filter_len*.



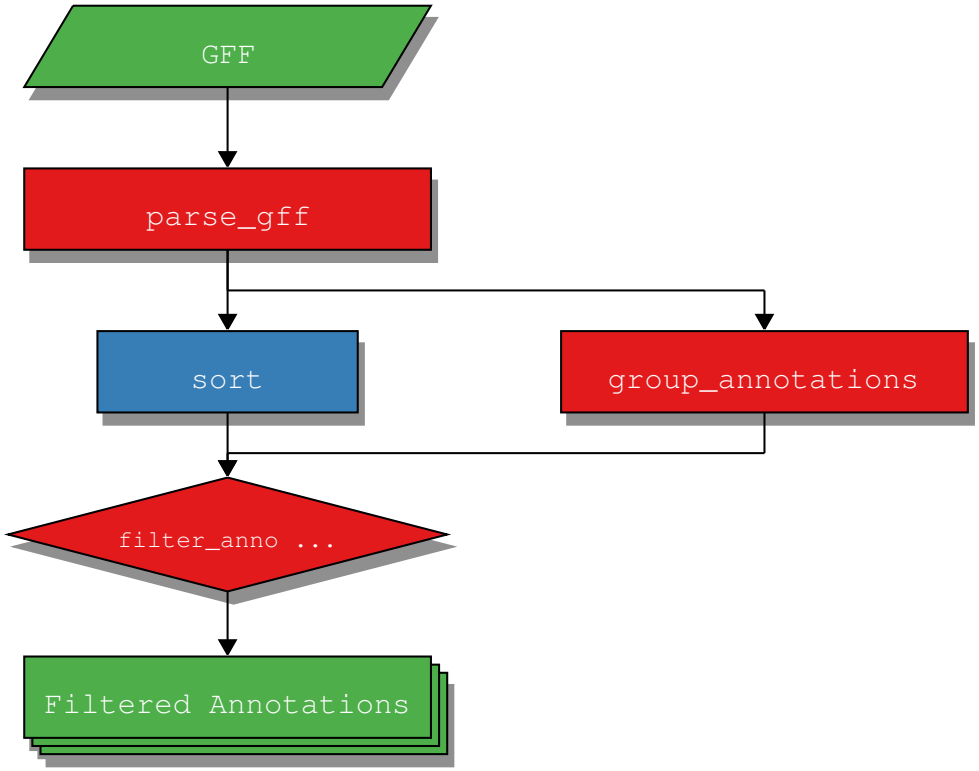
Overlap Filtering

Filters overlapping annotations using the functions `mgkit.filter.gff.choose_annotation()` and `mgkit.filter.gff.filter_annotations()`, after the annotations are grouped by both sequence and strand. If the GFF is sorted by sequence name and strand, the *-t* can be used to make the filtering use less memory. It can be sorted in Unix using *sort -s -k 1,1 -k 7,7 gff_file*, which applies a stable sort using the sequence name as the first key and the strand as the second key.

Note: It is also recommended to use:

```
export LC_ALL=C
```

To speed up the sorting



The above digram describes the internals of the script.

The annotations needs first to be grouped by seq_id and strand, forming a group that can be then be passed to *mgkit.filter.gff.filter_annotations()*. This function:

1. sort annotations by bit score, from the highest to the lowest
2. loop over all combination of N=2 annotations:
   (a) choose which of the two annotations to discard if they overlap for a the required amount of bp (defaults to 100bp)
   (b) in which case, the preference is given to the db quality first, than the bit score and finally the lenght of annotation, the one with the highest values is kept

While the default behaviour is the same, now it is posible to decided the function used to discard one the two annotations. It is possible to use the *-c* argument to pass a string that defines the function. The string passed must start with or without a **+**. Using **+** translates into the builtin function *max* while no **+** translates into *min* from the second character on, any number of attributes can be used, separated by commas. The attributes, however, must be one of the properties defined in *mgkit.io.gff.Annotation*, *bitscore* that returns the value converted in a *float*. Internally the attributes are stored as strings, so for attributes that have no properties in the class, such as *evalue*, the *float* builtin is applied.

The tuples built for both annotations are then passed to the comparison function to be selected and the value returned by it is **discarded**. The order of the elements in the string is important to define the priority given to each element in the comparison and the leftmost one has the highesst priority.

Examples of function strings:

- *-dbq,bitscore,length* becomes max((ann1.dbq, ann1.bitscore, ann1.length), (ann2.dbq, ann2.bitscore, ann2.length) - This is default and previously only choice
- *-bitscore,length,dbq* uses the same elements but gives lowest priority to *dbq*
- *+evalue*: will discard the annotation with the highest *evalue*

Per Sequence Values

The *sequence* command allows to filter on a per sequence basis, using functions such as the median, quantile and mean on attributes like evalue, bitscore and identity. The file can be passed as sorted already, saving memory (like in the *overlap* command), but it's not needed to sort the file by strand, only by the first column.

Coverage Filtering

The *cov* command calculates the coverage of annotations as a measure of the percentage of each reference sequence length. A minimum coverage percentage can be used to keep the annotations of sequences that have a greater or equal coverage than the specified one.

Changes

New in version 0.1.12.

Changed in version 0.1.13: added *–sorted* option

Changed in version 0.2.0: changed option *-c* to accept a string to filter overlap

Changed in version 0.2.5: added *sequence* command

Changed in version 0.2.6: added *length* as attribute and *min*/*max*, and *ge* is the default comparison for command *sequence*, *–sort-attr* to *overlap*

Changed in version 0.3.1: added *–num-gt* and *–num-lt* to *values* command, added *cov* command

4.2.2 Options

Filter GFF files

```
usage: filter-gff [-h] [-v | --quiet] [--cite] [--manual] [--version]
                  {values,overlap,sequence,cov} ...
```

Named Arguments

-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

Sub-commands:

values

Filter based on values

```
filter-gff values [-h] [-s SEQ_ID] [--strand {+,-}]
                  [-sl START_LOWER | -sg START_HIGHER]
                  [-el END_LOWER | -eg END_HIGHER]
                  [-lg LENGTH | -ls LENGTH_SHORT] [--source SOURCE]
                  [-f FEAT_TYPE] [-g GENE_ID] [-d DB] [-q DB_QUAL]
                  [-b BITSCORE] [-t TAXON_ID] [--str-eq STR_EQ]
                  [--str-in STR_IN] [--num-ge NUM_GE] [--num-le NUM_LE]
                  [--num-gt NUM_GT] [--num-lt NUM_LT] [-v | --quiet] [--cite]
                  [--manual] [--version]
                  [input_file] [output_file]
```

Positional Arguments

input_file	Input GFF file, defaults to stdin
	Default: -
output_file	Output GFF file, defaults to stdout
	Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-s, --seq-id	filter by sequence id
--strand	Possible choices: +, -
	filter by strand
-sl, --start-lower	returns only annotations where the start position is less than
-sg, --start-higher	**returns only annotations where the start position is greater** than
-el, --end-lower	**returns only annotations where the end position is equal or** less than
-eg, --end-higher	**returns only annotations where the end position is equal ot** greater than
-lg, --length	filter by annotation length equal to or longer than
-ls, --length-short	filter by annotation length equal to or shorter than
--source	filter by source
-f, --feat-type	filter by feature type
-g, --gene-id	filter by gene_id
-d, --db	filter by db
-q, --db-qual	filter by db quality equal or greater than
-b, --bitscore	filter by bitscore equal or greater than
-t, --taxon-id	filter by taxon_id
--str-eq	**filter by custom key:value, if the argument is 'key:value' the** annotation is kept if it contains an attribute 'key' whose value is exactly 'value' as a string.
--str-in	Same as '–str-eq' but 'value' is contained in the attribute

--num-ge	Same as '–str-eq' but 'value' is a number which is equal or greater than
--num-le	Same as '–num-ge' but 'value' is a number which is equal or less than
--num-gt	Same as '–str-eq' but 'value' is a number which is greater than
--num-lt	Same as '–num-ge' but 'value' is a number which is less than
-v, --verbose	more verbose - includes debug messages Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

overlap

Use overlapping filter

```
filter-gff overlap [-h] [-s SIZE] [-t] [-c CHOOSE_FUNC]
                   [-a {bitscore,identity,length}] [-v | --quiet] [--cite]
                   [--manual] [--version]
                   [input_file] [output_file]
```

Positional Arguments

input_file	Input GFF file, defaults to stdin Default: -
output_file	Output GFF file, defaults to stdout Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-s, --size	Size of the overlap that triggers the filter Default: 100
-t, --sorted	**If the GFF file is sorted (all of a sequence annotations are** contiguos and sorted by strand) can use less memory, *sort -s -k 1,1 -k 7,7* can be used Default: False
-c, --choose-func	Function to choose between two overlapping annotations Default: dbq,bitscore,length
-a, --sort-attr	Possible choices: bitscore, identity, length Attribute to sort annotations before filtering (default bitscore) Default: "bitscore"
-v, --verbose	more verbose - includes debug messages Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework

--manual	Show the script manual
--version	show program's version number and exit

sequence

Filter on a per sequence basis

```
filter-gff sequence [-h] [-t] [-a {evalue,bitscore,identity,length}]
                    [-m | -d | -q QUANTILE | -s STD | -x | -n]
                    [-c {gt,ge,lt,le}] [-v | --quiet] [--cite] [--manual]
                    [--version]
                    [input_file] [output_file]
```

Positional Arguments

input_file	Input GFF file, defaults to stdin
	Default: -
output_file	Output GFF file, defaults to stdout
	Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-t, --sorted	**If the GFF file is sorted (all of a sequence annotations are** contiguos) can use less memory, *sort -s -k 1,1* can be used
	Default: False
-a, --attribute	Possible choices: evalue, bitscore, identity, length
	Attribute on which to apply the filter
	Default: "bitscore"
-m, --mean	Filter by the mean value
	Default: False
-d, --median	Filter by the median value
	Default: False
-q, --quantile	Filter by the quantile value
-s, --std	**Filter by: mean + (X * std) where X is the number** supplied
-x, --max	Filter by the maximum value
	Default: False
-n, --min	Filter by the minimum value
	Default: False
-c, --comparison	Possible choices: gt, ge, lt, le
	Type of comparison (e.g. ge -> greater than or equal to)
	Default: "ge"
-v, --verbose	more verbose - includes debug messages
	Default: 20

--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

cov

Filter on a per coverage basis

```
filter-gff cov [-h] -f REFERENCE [-s] [-t] [-c MIN_COVERAGE] [-v | --quiet]
               [--cite] [--manual] [--version]
               [input_file] [output_file]
```

Positional Arguments

input_file	Input GFF file, defaults to stdin
	Default: -
output_file	Output GFF file, defaults to stdout
	Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-f, --reference	Reference FASTA file for the GFF
-s, --strand-specific	If the coverage must be calculated on each strand
	Default: False
-t, --sorted	Assumes the GFF to be correctly sorted
	Default: False
-c, --min-coverage	Minimum coverage for the contig/strand
	Default: 0.0
-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

4.3 add-gff-info - Add informations to GFF annotations

4.3.1 Overview

Add more information to GFF annotations: gene mappings, coverage, taxonomy, etc..

Uniprot Command

If the *gene_id* of an annotation is a Uniprot ID, the script queries Uniprot for the requested information. At the moment the information that can be added is the taxon_id, taxon_name, lineage and mapping to EC, KO, eggNOG IDs.

It's also possible to add mappings to other databases using the *-m* option with the correct identifier for the mapping, which can be found at this page[56]; for example if it's we want to add the mappings of uniprot IDs to *BioCyc*, in the *abbreviation* column of the mappings we find that it's identifier is *REACTOME_ID*, so we pass *-m REACTOME* to the script (leaving *_ID* out). Mapped IDs are separated by commas.

The taxonomy IDs are not overwritten if they are found in the annotations, the *-f* is provided to force the overwriting of those values.

See also *MGKit GFF Specifications* for more informations about the GFF specifications used.

Note: As the script needs to query Uniprot a lot, it is recommended to split the GFF in several files, so an error in the connection doesn't waste time.

However, a cache is kept to reduce the number of connections

Taxonomy Command

To refine the taxonomic assignments of predicted genes annotations, the annotation sequences may be searched against a database like the NCBI *nt*.

This commands takes as input a GFF file, one or more blast output files and a file with all mappings from GIDs to taxonomy IDs. More information on how to get the file can be read in the documentation of the function `mgkit.io.blast.parse_gi_taxa_table()`.

The fasta sequences used with BLAST must have as name the uid of the annotations they refer to, and one way to obtain these sequences is to use the function *`mgkit.io.gff.extract_nuc_seqs()`* and save them to a fasta file. Another options is to use the *sequence* command of the *get-gff-info* script (*get-gff-info - Extract informations to GFF annotations*).

The command accept a minimum bitscore to accept an hit and the taxon ID is selected by default using top hit method, but LCA can be used, using the *-l* switch.

Top Hit

The best hit is selected from all those found for a sequence which has the maximum bitscore and identity, with the bitscore having the highest priority.

LCA Taxon

Activated with the *-l* switch, it selects the last common ancestor of all taxon IDs that are from the cellular organism root in the taxonomy and are within a 10 bits (by default, can be customised with *-a*) from the hit with the highest bitscore. If a taxon ID is not found in the taxonomy, it is excluded. One of the requirements of this option is a file that contains the full taxonomy from Uniprot/NCBI. The file can be obtained with the following command:

```
$ download_data -x -p -m your@email
```

The command will output a *taxonomy.pickle* file that can be passed to the *-x* option *download-data - Download Taxonomy from NCBI*.

[56] http://www.uniprot.org/faq/28

Coverage Command

Adds coverage information from BAM alignment files to a GFF file, using the function *mgkit.align.add_coverage_info()*, the user needs to supply for each sample a BAM file, using the *-a* option, whose parameter is in the form *sample,samplealg.bam*. More samples can be supplied adding more *-a* arguments.

Hint: As an example, to add coverage for *sample1*, *sample2* the command line is:

```
add-gff-info coverage -a sample1,sample1.bam -a sample2,sample2.bam \
inputgff outputgff
```

A total coverage for the annotation is also calculated and stored in the *cov* attribute, while each sample coverage is stored into *sample_cov* as per *MGKit GFF Specifications*.

Adding Coverage from samtools depth

The *cov_samtools* allows the use of the output of *samtools* **depth** command. The *-aa* options must be used to pass information about all base pairs and sequences coverage in the BAM/SAM file. The command accept only one sample and the relative file/stream from *samtools*, meaning that multiple samples coverage information must be added one at a time. One solution is to pipe multiple commands to obtain result wanted. For example:

```
$ add-gff-info cov_samtools -s SAMPLE1 -d sample1-coverage input.gff |          add-
↪gff-info cov_samtools -s SAMPLE2 -d sample2-coverage - output.gff
```

This command will add the coverage information for SAMPLE1 and SAMPLE2 from the respective files.

Uniprot Offline Mappings

Similar to the *uniprot* command, it uses the idmapping[57] file provided by Uniprot, which speeds up the process of adding mappings and taxonomy IDs from Uniprot gene IDs. It's not possible tough to add *EC* mappings with this command, as those are not included in the file.

Kegg Information

The *kegg* command allows to add information to each annotation. Right now the information that can be added is restricted to the pathway(s) (reference KO) a KO is part of and both the KO and pathway(s) descriptions. This information is stored in keys starting with **ko_**.

Expected Aminoacidic Changes

Some scripts, like *snp_parser - SNPs analysis*, require information about the expected number of synonymous and non-synonymous changes of an annotation. This can be done using *mgkit.io.gff.Annotation.add_exp_syn_count()* by the user of the command *exp_syn* of this script. The attributes added to each annotation are explained in the *MGKit GFF Specifications*

Adding Information from eggNOG

The *eggnog* command allows to add information from the *annotations* file available for profiles in eggNOG.

[57] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping.dat.gz

Adding Count Data

Count data on a per-sample basis can be added with the *counts* command. The accepted inputs are from HTSeq-count and featureCounts. The ouput produced by featureCounts, is the one from using its **-f** option must be used.

This script accept by default a tab separated file, with a uid in the first column and the other columns are the counts for each sample, in the same order as they are passed to the **-s** option. To use the featureCounts file format, this script **-e** option must be used.

The sample names must be provided in the same order as the columns in the input files. If the counts are FPKMS the *-f* option can be used.

Adding Taxonomy from a Table

There are cases where it may needed or preferred to add the taxonomy from a *gene_id* already provided in the GFF file. For such cases the *addtaxa* command can be used. It works in a similar way to the *taxonomy* command, only it expect three different type of inputs:

- *GI-Taxa* table from NCBI (e.g. gi_taxid_nucl.dmp,)
- tab separated table
- dictionary
- HDF5

The first two are tab separated files, where on each line, the first column is the *gene_id* that is found in the first column, while the second if the *taxon_id*.

The third option is a serialised Python *dict*/hash table, whose keys are the *gene_id* and the value is that gene corresponding *taxon_id*. The serialised formats accepted are msgpack, json and pickle. The *msgpack* module must be importable. The option to use json and msgpack allow to integrate this script with other languages without resorting to a text file.

The last option is a HDF5 created using the *to_hdf* command in *taxon-utils - Taxonomy Utilities*. This requires *pandas* installed and *pytables* and it provides faster lookup of IDs in the table.

While the default is to look for the *gene_id* attribute in the GFF annotation, another attribute can be specified, using the **-gene-attr** option.

Note: the dictionary content is loaded after the table files and its keys and corresponding values takes precedence over the text files.

Warning: from September 2016 NCBI will retire the GI. In that case the same kind of table can be built from the *nucl_gb.accession2taxid.gz* file The format is different, but some information can be found in `mgkit.io.blast.parse_accession_taxa_table()`

Adding information from Pfam

Adds the Pfam description for the annotation, by downloading the list from Pfam.

The options allow to specify in which attribute the ID/ACCESSION is stored (defaults to *gene_id*) and which one between ID/ACCESSION is the value of that attribute (defaults to *ID*). if no description is found for the family, a warning message is logged.

Changes

Changed in version 0.3.0: added *cov_samtools* command, *–split* option to *exp_syn*, *-c* option to *addtaxa*

Changed in version 0.2.6: added *skip-no-taxon* option to *addtaxa*

Changed in version 0.2.5: if a dictionary is supplied to *addtaxa*, the GFF is not preloaded

Changed in version 0.2.3: added *pfam* command, renamed *gitaxa* to *addtaxa* and made it general

Changed in version 0.2.2: added *eggnog*, *gitaxa* and *counts* command

Changed in version 0.2.1.

- added *-d* to *uniprot* command
- added cache to *uniprot* command
- added *kegg* command (cached)

Changed in version 0.1.16: added *exp_syn* command

Changed in version 0.1.15: *taxonomy* command *-b* option changed

Changed in version 0.1.13.

- added *–force-taxon-id* option to the *uniprot* command
- added *coverage* command
- added *taxonomy* command
- added *unipfile* command

New in version 0.1.12.

4.3.2 Options

Adds informations to a GFF file

```
usage: add-gff-info [-h] [-v | --quiet] [--cite] [--manual] [--version]
                    {uniprot,taxonomy,coverage,exp_syn,unipfile,kegg,eggnog,counts,
↪addtaxa,pfam,cov_samtools}
                    ...
```

Named Arguments

-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

Sub-commands:

uniprot

Adds information from GFF whose gene_id is from Uniprot

```
add-gff-info uniprot [-h] [-c EMAIL] [--buffer BUFFER] [-f] [-t] [-l] [-e]
                     [-ec] [-ko] [-d] [-m MAPPING] [-v | --quiet] [--cite]
                     [--manual] [--version]
                     [input_file] [output_file]
```

Positional Arguments

input_file	Input GFF file, defaults to stdin
	Default: -
output_file	Output GFF file, defaults to stdout
	Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-c, --email	Contact email
--buffer	Number of annotations to keep in memory
	Default: 50
-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

Requires Internet connection

-f, --force-taxon-id	Overwrite taxon_id if already present
	Default: False
-t, --taxon-id	Add taxonomic ids to annotations, if taxon_id is found, it won't be Overwritten.
	Default: False
-l, --lineage	Add taxonomic lineage to annotations
	Default: False
-e, --eggnog	Add eggNOG mappings to annotations
	Default: False
-ec	Add EC mappings to annotations
	Default: False
-ko	Add KO mappings to annotations
	Default: False
-d, --protein-names	Add Uniprot description
	Default: False
-m, --mapping	Add any DB mappings to annotations

taxonomy

Adds taxonomic information from annotation sequences blasted against a NCBI db

```
add-gff-info taxonomy [-h] -t GI_TAXA_TABLE -b BLAST_OUTPUT [-s BITSCORE]
                      [-d TAXON_DB] [-l] [-x TAXONOMY] [-a MAX_DIFF]
                      [-v | --quiet] [--cite] [--manual] [--version]
                      [input_file] [output_file]
```

Positional Arguments

input_file	Input GFF file, defaults to stdin
	Default: -
output_file	Output GFF file, defaults to stdout
	Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-t, --gi-taxa-table	GIDs taxonomy table (e.g. gi_taxid_nucl.dmp.gz)
-b, --blast-output	BLAST output file(s)
-s, --bitscore	Minimum bitscore allowed
	Default: 40
-d, --taxon-db	NCBI database used
	Default: "NCBI-NT"
-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

LCA options

-l, --lca	Use last common ancestor to solve ambiguities
	Default: False
-x, --taxonomy	Taxonomy file
-a, --max-diff	Bitscore difference from the max hit
	Default: 10

coverage

Adds coverage information from BAM Alignment files

```
add-gff-info coverage [-h] -a SAMPLE_ALIGNMENT [-v | --quiet] [--cite]
                      [--manual] [--version]
                      [input_file] [output_file]
```

Positional Arguments

input_file Input GFF file, defaults to stdin

Default: -

output_file Output GFF file, defaults to stdout

Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-a, --sample-alignment sample name and correspondent alignment file separated by comma

-v, --verbose more verbose - includes debug messages

Default: 20

--quiet less verbose - only error and critical messages

--cite Show citation for the framework

--manual Show the script manual

--version show program's version number and exit

exp_syn

Adds expected synonymous and non-synonymous changes information

```
add-gff-info exp_syn [-h] -r REFERENCE [-s] [-v | --quiet] [--cite] [--manual]
                     [--version]
                     [input_file] [output_file]
```

Positional Arguments

input_file Input GFF file, defaults to stdin

Default: -

output_file Output GFF file, defaults to stdout

Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-r, --reference reference sequence in fasta format

-s, --split **Split the sequence header of the reference at the first** space, to emulate BLAST behaviour

Default: False

-v, --verbose more verbose - includes debug messages

Default: 20

--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

unipfile

Adds mappings and taxonomy from Uniprot mapping file

```
add-gff-info unipfile [-h] -i MAPPING_FILE [-f] -m
                      {EMBL-CDS,KEGG,eggNOG,EMBL,STRING,UniPathway,BioCyc,NCBI_
↪TaxID,KO,GI}
                      [-v | --quiet] [--cite] [--manual] [--version]
                      [input_file] [output_file]
```

Positional Arguments

input_file	Input GFF file, defaults to stdin
	Default: -
output_file	Output GFF file, defaults to stdout
	Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-i, --mapping-file	Uniprot mapping file
	Default: "idmapping.dat.gz"
-f, --force-taxon-id	Overwrite taxon_id if already present
	Default: False
-m, --mapping	Possible choices: EMBL-CDS, KEGG, eggNOG, EMBL, STRING, UniPathway, BioCyc, NCBI_TaxID, KO, GI
	Mappings to add
-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

kegg

Adds information and mapping from Kegg

```
add-gff-info kegg [-h] [-c EMAIL] [-d] [-p] [-m KEGG_ID] [-v | --quiet]
                  [--cite] [--manual] [--version]
                  [input_file] [output_file]
```

Positional Arguments

input_file Input GFF file, defaults to stdin

Default: -

output_file Output GFF file, defaults to stdout

Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-c, --email Contact email

-v, --verbose more verbose - includes debug messages

Default: 20

--quiet less verbose - only error and critical messages

--cite Show citation for the framework

--manual Show the script manual

--version show program's version number and exit

Requires Internet connection

-d, --description Add Kegg description

Default: False

-p, --pathways Add pathways ID involved

Default: False

-m, --kegg-id In which attribute the Kegg ID is stored (defaults to *gene_id*)

Default: "gene_id"

eggnog

Adds information from eggNOG

```
add-gff-info eggnog [-h] -a ANNOTATIONS_FILE [-v | --quiet] [--cite]
                    [--manual] [--version]
                    [input_file] [output_file]
```

Positional Arguments

input_file Input GFF file, defaults to stdin

Default: -

output_file Output GFF file, defaults to stdout

Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-a, --annotations-file	Annotations file
-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

counts

Adds counts data to the GFF

```
add-gff-info counts [-h] -s SAMPLES -c COUNT_FILES [-f] [-e] [-v | --quiet]
                    [--cite] [--manual] [--version]
                    [input_file] [output_file]
```

Positional Arguments

input_file	Input GFF file, defaults to stdin
	Default: -
output_file	Output GFF file, defaults to stdout
	Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-s, --samples	Comma separated sample names, in the same order as the count file
-c, --count-files	Count file(s)
-f, --fpkms	If the counts are FPKMS
	Default: False
-e, --featureCounts	If the counts files are from featureCounts (using the -f option)
	Default: False
-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

addtaxa

Adds taxonomy information from a GI-Taxa, gene_id/taxon_id table or a dictionary serialised as a pickle/msgpack/json file

```
add-gff-info addtaxa [-h] [-t GENE_TAXON_TABLE] [-f HDF_TABLE] [-a GENE_ATTR]
                     [-x TAXONOMY] [-d DICTIONARY] [-e] [-db TAXON_DB] [-c]
                     [-v | --quiet] [--cite] [--manual] [--version]
                     [input_file] [output_file]
```

Positional Arguments

input_file Input GFF file, defaults to stdin

Default: -

output_file Output GFF file, defaults to stdout

Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-t, --gene-taxon-table **GIDs taxonomy table (e.g. gi_taxid_nucl.dmp.gz) or a similar** file where GENE/TAXON are tab separated and one per line

-f, --hdf-table HDF5 file and table name to use for taxon_id lookups. The format to pass is the file name, colon and the table file.hf5:taxa-table. The index in the table is the accession_id, while the column *taxon_id* stores the taxon_id as int

-a, --gene-attr In which attribute the GENEID in the table is stored (defaults to *gene_id*)

Default: "gene_id"

-x, --taxonomy Taxonomy file - If given, both *taxon_name* and *lineage* attributes will be set

-d, --dictionary A serialised dictionary, where the key is the GENEID and the value is TAXONID. It can be in json or msgpack format (can be a compressed file) *Note*: the dictionary values takes precedence over the table files

-e, --skip-no-taxon If used, annotations with no taxon_id won't be outputted

Default: False

-db, --taxon-db DB used to add the taxonomic information

Default: "NONE"

-c, --cache-table **If used, annotations are not preloaded, but the taxa table is** cached, instead.

Default: False

-v, --verbose more verbose - includes debug messages

Default: 20

--quiet less verbose - only error and critical messages

--cite Show citation for the framework

--manual Show the script manual

--version show program's version number and exit

pfam

Adds information from Pfam

```
add-gff-info pfam [-h] [-i ID_ATTR] [-a] [-v | --quiet] [--cite] [--manual]
                  [--version]
                  [input_file] [output_file]
```

Positional Arguments

input_file Input GFF file, defaults to stdin

Default: -

output_file Output GFF file, defaults to stdout

Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-i, --id-attr In which attribute the Pfam ID/ACCESSION is stored (defaults to *gene_id*)

Default: "gene_id"

-a, --use-accession **If used, the attribute value is the Pfam ACCESSION** (e.g. PF06894), not ID (e.g. Phage_TAC_2)

Default: False

-v, --verbose more verbose - includes debug messages

Default: 20

--quiet less verbose - only error and critical messages

--cite Show citation for the framework

--manual Show the script manual

--version show program's version number and exit

cov_samtools

Adds information from samtools_depth

```
add-gff-info cov_samtools [-h] [-s SAMPLE] -d DEPTH [-n NUM_SEQS]
                          [-v | --quiet] [--cite] [--manual] [--version]
                          [input_file] [output_file]
```

Positional Arguments

input_file Input GFF file, defaults to stdin

Default: -

output_file Output GFF file, defaults to stdout

Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-s, --sample	sample name
-d, --depth	*samtools depth -aa* file
-n, --num-seqs	Number of sequences to update the log
	Default: 10000
-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

4.4 get-gff-info - Extract informations to GFF annotations

4.4.1 Overview

4.4.2 Options

4.5 hmmer2gff - Convert HMMER output to GFF

4.5.1 Overview

Script to convert HMMER results files (domain table) to a GFF file, the name of the profiles are expected to be now in the form *GENEID_TAXONID_TAXON-NAME(-nr)* by default, but any other profile name is accepted.

The profiles tested are those made from Kegg Orthologs, from the *download_profiles* script. If the *–no-custom-profiles* options is used, the script can be used with any profile name. The profile name will be used for *gene_id*, *taxon_id* and *taxon_name* in the GFF file.

It is possible to use seuqnces not translated using mgkit, no information on the frame is assumed, so this script can be used against a protein DB. For example Uniprot can be searched for profiles, in which case the **–no-frame** options must be used.

Note: for GENEID, old documentation points to KOID, it is the same

---

Warning: The compatibility with old data has been **removed**, meaning that old experiments must use the scripts from those versions. It is possible to use multiple environments, with *virtualenv* for this purpose. An examples is given in *Installation*.

Changes

Changed in version 0.1.15: adapted to new GFF module and specs

Changed in version 0.2.1: added options to customise output and filters and old restrictions

Changed in version 0.3.1: added *–no-frame* option for non mgkit-translated proteins, sequence headers are handled the same way as HMMER (truncated at the first space)

4.5.2 Options

Convert HMMER data to GFF file

```
usage: hmmer2gff [-h] [-o [OUTPUT_FILE]] [-t DISCARD] [-d] [-c] [-db DATABASE]
                 [-f FEATURE_TYPE] [-n] [-v | --quiet] [--cite] [--manual]
                 [--version]
                 aa_file [hmmer_file]
```

Named Arguments

-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

File options

aa_file	Fasta file containing contigs translated to aa (used by HMMER)
hmmer_file	Default: -
-o, --output-file	Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Filters

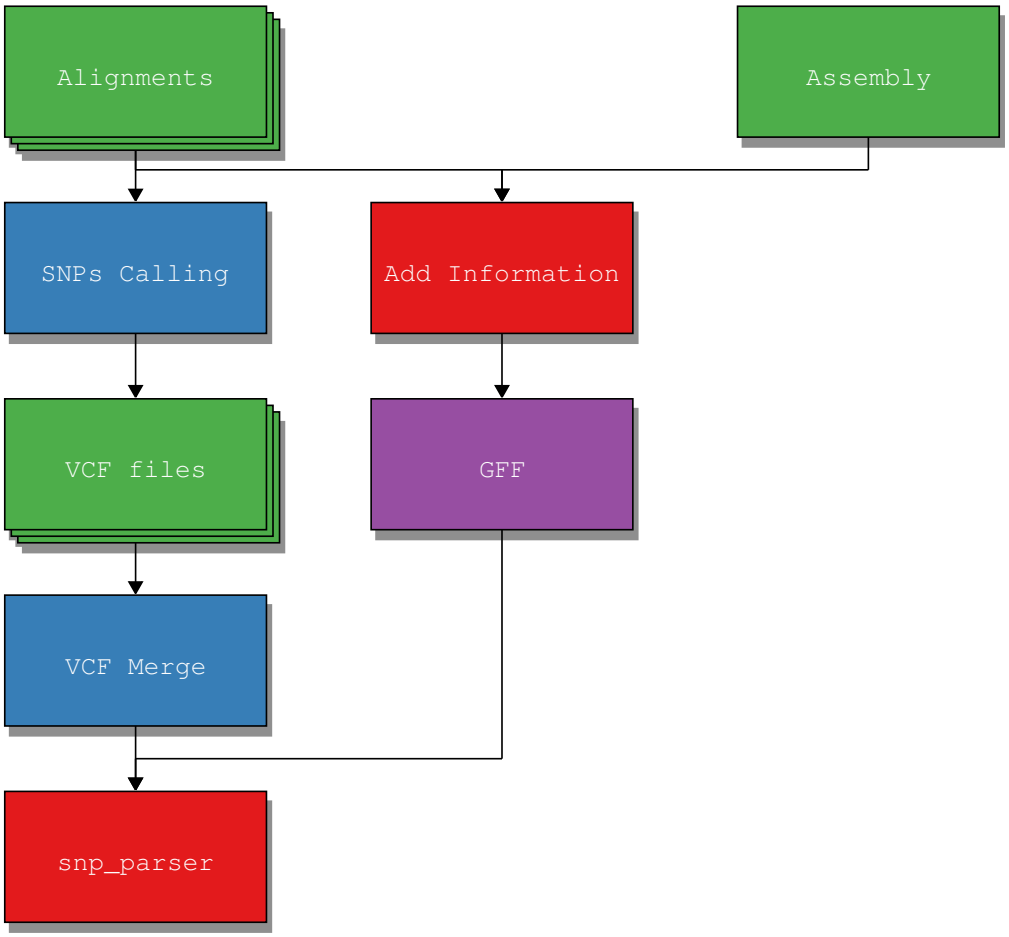
-t, --discard	Evalue over which an hit will be discarded
	Default: 0.05
-d, --disable-evalue	Disable Evalue filter
	Default: False

GFF

-c, --no-custom-profiles	Profiles names are not in the custom format
	Default: True
-db, --database	Database from which the profiles are generated " +" (e.g. PFAM)
	Default: "CUSTOM"
-f, --feature-type	Type of feature (e.g. gene)
	Default: "gene"
-n, --no-frame	Set if the sequences were not translated with translate_seq
	Default: False

4.6 snp_parser - SNPs analysis

4.6.1 Overview



The workflow starts with a number of alignments passed to the SNP calling software, which produces one VCF file per alignment/sample. These VCF files are used by SNPDat[58] along a GTF file and the reference genome to integrate the information in VCF files with synonymous/non-synonymous information.

All VCF files are merged into a VCF that includes information about all the SNPs called among all samples. This merged VCF is passed, along with the results from SNPDat and the GFF file to snp_parser.py which integrates information from all data sources and output files in a format that can be later used by the rest of the pipeline.[59]

Note: The GFF file passed to the parser must have per sample coverage information.

[58] http://code.google.com/p/snpdat/
[59] This step is done separately because it's both time consuming and can helps to paralellise later steps

4.6.2 Script Reference

4.6.3 Options

4.7 download-taxonomy.sh Download Taxonomy

A bash script called **download-taxonomy.sh** is installed along with MGKit. This script download the relevant files from NCBI using *wget*, and save the taxonomy file that can be used with MGKit to a file called **taxonomy.pickle**.

Since the script uses *wget* to download the file taxdump.tar.gz[60], if *wget* can't be found, the scripts fails. To avoid this situation, the file can be downloaded in another way, and the script detects if the file exists, avoiding the call of *wget*.

The script can also save the file with another file name, if this is passed when the script is invoked. if the file extension contains *.msgpack*, the **msgpack** module is used to write the taxonomy, otherwise *pickle* is used.

The advantage of *msgpack* is faster read/write and better compression ratio; it needs an additional module (msgpack[61]) that is not installed by default.

4.8 taxon-utils - Taxonomy Utilities

4.8.1 Overview

The script contains commands used to access functionality related to taxonomy, without the need to write ad-hoc code for functionality that can be part of a workflow. One example is access to the the last common ancestor function contained in the *mgkit.taxon*.

Last Common Ancestor (lca and lca_line)

These commands expose the functionality of `last_common_ancestor_multiple()`, making it accessible via the command line. They differ in the input file format and the choice of output files.

the *lca* command can be used to define the last common ancestor of contigs from the annotation in a GFF file. The command uses the *taxon_ids* from all annotations belonging to a contig/sequence, if they have a **bitscore** higher or equal to the one passed (50 by default). The default output of the command is a tab separated file where the first column is the contig/sequence name, the taxon_id of the last common ancestor, its scientific/common name and its lineage.

For example:

```
contig_21   172788  uncultured phototrophic eukaryote   cellular organisms,
↪environmental samples
```

If the *-r* is used, by passing the fasta file containing the nucleotide sequences the output file is a GFF where for each an annotation for the full contig length contains the same information of the tab separated file format.

The **lca_line** command accept as input a file where each line consist of a list of taxon_ids. The separator for the list can be changed and it defaults to TAB. The last common ancestor for all taxa on a line is searched. The ouput of this command is the same as the tab separated file of the **lca** command, with the difference that instead of the first column, which in this command becames a list of all *taxon_ids* that were used to find the last common ancestor for that line. The list of *taxon_ids* is separated by semicolon ";".

Note: Both also accept the *-n* option, to report the config/line and the taxon_ids that had no common ancestors. These are treated as errors and do not appear in the output file.

---

[60] ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdump.tar.gz
[61] https://github.com/msgpack/msgpack-python

Krona Output

New in version 0.3.0.

The *lca* command supports the writing of a file compatible with Krona. The output file can be used with the *ktImportText/ImportText.pl* script included with KronaTools[62]. Specifically, the output from *taxon_utils* will be a file with all the lineages found (tab separated), that can be used with:

```
$ ktImportText -q taxon_utils_ouput
```

Note the use of *-q* to make the script count the lineages. Sequences with no LCA found will be marked as *No LCA* in the graph, the *-n* is not required.

Note: Please note that the output won't include any sequence that didn't have a hit with the software used. If that's important, the **-kt** option can be used to add a number of *Unknown* lines at the end, to read the total supplied.

Filter by Taxon

The **filter** command of this script allows to filter a GFF file using the *taxon_id* attribute to include only some annotations, or exclude some. The filter is based on the *mgkit.taxon.is_ancestor* function, and the *mgkit.filter.taxon.filter_taxon_by_id_list*. It allows to pass a list of taxon_id (or taxon_names) to the script. The *include* filter will only output annotations that have one of the passed taxa as ancestors, while the *exclude* filter will remove those annotations, that have the passed taxa as ancestors, from the output.

A list of comma separated taxon_ids can be supplied, as for the names. If any of the the supplied names have multiple taxon_id (e.g. Actinobacteria) the script exits and in the log can be found the list of duplicates. For cases like this, it's preferred for the user to supply a taxon_id, as they can be searched in NCBI taxonomy (also Uniprot).

Warning: Annotations with no taxon_id are not included in the output of both filters

Convert Taxa Tables to HDF5

This command is used to convert the taxa tables download from Uniprot and NCBI, using the scripts mentioned in *download-data - Download Taxonomy from NCBI*, *download-uniprot-taxa.sh* and *download-ncbi-taxa* into a HDF5 file that can be used with the *addtaxa* command in *add-gff-info - Add informations to GFF annotations*.

The advantage is a faster lookup of the IDs. The other is a smaller memory footprint when a great number of annotations are kept in memory.

Changes

Changed in version 0.3.1: added *to_hdf* command

Changed in version 0.3.1: added *-j* option to *lca*, which outputs a JSON file with the LCA results

Changed in version 0.3.0: added *-k* and *-kt* options for Krona output, lineage now includes the LCA also added *-a* option to select between lineages with only ranked taxa. Now it defaults to all components.

Changed in version 0.2.6: added *feat-type* option to *lca* command, added phylum output to nolca

New in version 0.2.5.

[62] https://github.com/marbl/Krona/wiki

4.8.2 Options

Taxonomy Utilities

```
usage: taxon_utils [-h] [-v | --quiet] [--cite] [--manual] [--version]
                   {lca,lca_line,filter,to_hdf} ...
```

Named Arguments

-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

Sub-commands:

lca

Finds the last common ancestor for each sequence in a GFF file

```
taxon_utils lca [-h] [-b BITSCORE] [-s] [-a] [-ft FEAT_TYPE]
                [-r REFERENCE | -k | -j] [-kt KRONA_TOTAL] [-n NO_LCA] -t
                TAXONOMY [-v | --quiet] [--cite] [--manual] [--version]
                [input_file] [output_file]
```

Positional Arguments

input_file	Input file, defaults to stdin
	Default: -
output_file	Output file, defaults to stdout
	Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-b, --bitscore	Minimum bitscore accepted
	Default: 0
-s, --sorted	**If the GFF file is sorted (all of a sequence annotations are** contiguos) can use less memory, *sort -s -k 1,1* can be used
	Default: False
-a, --only-ranked	**If set, only taxa that have a rank will be used in the lineage.** This is not advised for lineages such as Viruses, where the top levels have no rank
	Default: False
-ft, --feat-type	Feature type used if the output is a GFF (default is *LCA*)
	Default: "LCA"

-r, --reference	Reference file for the GFF, if supplied a GFF file is the output
-k, --krona	Output a file that can be read by Krona (text) Default: False
-j, --json	If used, the output is a JSON file with the LCA information Default: False
-kt, --krona-total	**Total number of raw sequences (used to output correct** percentages in Krona
-n, --no-lca	File to which write records with no LCA
-t, --taxonomy	Taxonomy file
-v, --verbose	more verbose - includes debug messages Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

lca_line

Finds the last common ancestor for all IDs in a text file line

```
taxon_utils lca_line [-h] [-s SEPARATOR] [-n NO_LCA] -t TAXONOMY
                     [-v | --quiet] [--cite] [--manual] [--version]
                     [input_file] [output_file]
```

Positional Arguments

input_file	Input file, defaults to stdin Default: -
output_file	Output file, defaults to stdout Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-s, --separator	separator for taxon_ids (defaults to TAB) Default: " "
-n, --no-lca	File to which write records with no LCA
-t, --taxonomy	Taxonomy file
-v, --verbose	more verbose - includes debug messages Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

filter

Filter a GFF file based on taxonomy

```
taxon_utils filter [-h]
                   [-i INCLUDE_TAXON_ID | -in INCLUDE_TAXON_NAME | -e EXCLUDE_
↪TAXON_ID | -en EXCLUDE_TAXON_NAME]
                   -t TAXONOMY [-v | --quiet] [--cite] [--manual] [--version]
                   [input_file] [output_file]
```

Positional Arguments

input_file Input file, defaults to stdin

Default: -

output_file Output file, defaults to stdout

Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-i, --include-taxon-id Include only taxon_ids (comma separated)

-in, --include-taxon-name Include only taxon_names (comma separated)

-e, --exclude-taxon-id Exclude taxon_ids (comma separated)

-en, --exclude-taxon-name Exclude taxon_names (comma separated)

-t, --taxonomy Taxonomy file

-v, --verbose more verbose - includes debug messages

Default: 20

--quiet less verbose - only error and critical messages

--cite Show citation for the framework

--manual Show the script manual

--version show program's version number and exit

to_hdf

Convert a taxa table to HDF5

```
taxon_utils to_hdf [-h] [-n TABLE_NAME] [-w] [-s INDEX_SIZE] [-c CHUNK_SIZE]
                   [-v | --quiet] [--cite] [--manual] [--version]
                   [input_file] [output_file]
```

Positional Arguments

input_file Input file, defaults to stdin

Default: -

output_file Output file, defaults to (taxa-table.hf5)

Default: "taxa-table.hf5"

Named Arguments

-n, --table-name	Name of the table/storage to use Default: "taxa"
-w, --overwrite	Overwrite the file, instead of appending to it Default: False
-s, --index-size	Maximum number of characters for the gene_id Default: 12
-c, --chunk-size	Chunk size to use when reading the input file Default: 5000000
-v, --verbose	more verbose - includes debug messages Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

4.9 fasta-utils - Fasta Utilities

4.9.1 Overview

New in version 0.3.0.

Scripts that includes some functionality to help use FASTA files with the framework

split command

Used to split a fasta file into smaller fragments

translate command

Used to translate nucleotide sequences into amino acids.

uid command

Used to change a FASTA file headers to a unique ID. A table (tab separated) with the changes made can be kept, using the *–table* option.

Changes

New in version 0.3.0.

Changed in version 0.3.1: added *translate* and *uid* command

4.9.2 Options

FASTA utilities

```
usage: fasta-utils [-h] [-v | --quiet] [--cite] [--manual] [--version]
                   {split,translate,uid} ...
```

Named Arguments

-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

Sub-commands:

split

Splits a FASTA file in a number of fragments

```
fasta-utils split [-h] [-p PREFIX] [-n NUMBER] [-z] [-v | --quiet] [--cite]
                  [--manual] [--version]
                  [input_file]
```

Positional Arguments

input_file	Input FASTA file, defaults to stdin
	Default: -

Named Arguments

-p, --prefix	Prefix for the file name in output
	Default: "split"
-n, --number	Number of chunks into which split the FASTA file
	Default: 10
-z, --gzip	gzip output files
	Default: False
-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

translate

Translate FASTA file in all 6 frames

```
fasta-utils translate [-h]
                      [-t {bac_plt,drs_mit,inv_mit,prt_mit,universal,vt_mit,yst_
↪alt,yst_mit}]
                      [-v | --quiet] [--cite] [--manual] [--version]
                      [input_file] [output_file]
```

Positional Arguments

input_file Input FASTA file, defaults to stdin

Default: -

output_file Input FASTA file, defaults to stdin

Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-t, --trans-table Possible choices: bac_plt, drs_mit, inv_mit, prt_mit, universal, vt_mit, yst_alt, yst_mit

translation table

Default: "universal"

-v, --verbose more verbose - includes debug messages

Default: 20

--quiet less verbose - only error and critical messages

--cite Show citation for the framework

--manual Show the script manual

--version show program's version number and exit

uid

Changes the header to a uid (unique ID)

```
fasta-utils uid [-h] [-t TABLE] [-v | --quiet] [--cite] [--manual] [--version]
                [input_file] [output_file]
```

Positional Arguments

input_file Input FASTA file, defaults to stdin

Default: -

output_file Input FASTA file, defaults to stdin

Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-t, --table	Filename of table of the changes (by default discards it)
-v, --verbose	more verbose - includes debug messages Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

4.10 fastq-utils - Fastq Utilities

4.10.1 Overview

4.10.2 Options

4.11 json2gff - Convert JSON to GFF

4.11.1 Overview

New in version 0.2.6.

This script converts annotations in JSON format that were created using MGKit back into GFF annotations.

mongodb command

Annotations converted into MongoDB records with *get-gff-info mongodb* can be converted back into a GFF file using this command. It can be useful to get a GFF file as output from a query to a MongoDB instance on the command line.

For example:

mongoexport -d db -c test | json2gff mongodb

will convert all the annotations in the database *db*, collection *test* to the standard out.

4.11.2 Options

Convert JSON to GFF

```
usage: json2gff [-h] [-v | --quiet] [--cite] [--manual] [--version]
                {mongodb,json} ...
```

Named Arguments

-v, --verbose	more verbose - includes debug messages Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework

--manual	Show the script manual
--version	show program's version number and exit

Sub-commands:

mongodb

Convert annotations from a MongoDB instance to GFF

```
json2gff mongodb [-h] [-v | --quiet] [--cite] [--manual] [--version]
                 [input_file] [output_file]
```

Positional Arguments

input_file	Input file, defaults to stdin
	Default: -
output_file	Output file, defaults to stdout
	Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

json

Convert annotations from JSON lines to GFF

```
json2gff json [-h] [-v | --quiet] [--cite] [--manual] [--version]
              [input_file] [output_file]
```

Positional Arguments

input_file	Input file, defaults to stdin
	Default: -
output_file	Output file, defaults to stdout
	Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-v, --verbose	more verbose - includes debug messages Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

4.12 download-profiles - Download Custom Profiles

4.12.1 Overview

This script downloads sequence data for each gene of interest (ortholog) and all the specified taxa. The files that are downloaded with this script can then be used to create HMMER profiles, to search for similarity in a aminoacidic or nucleotidic sequence.

4.12.2 Limitations

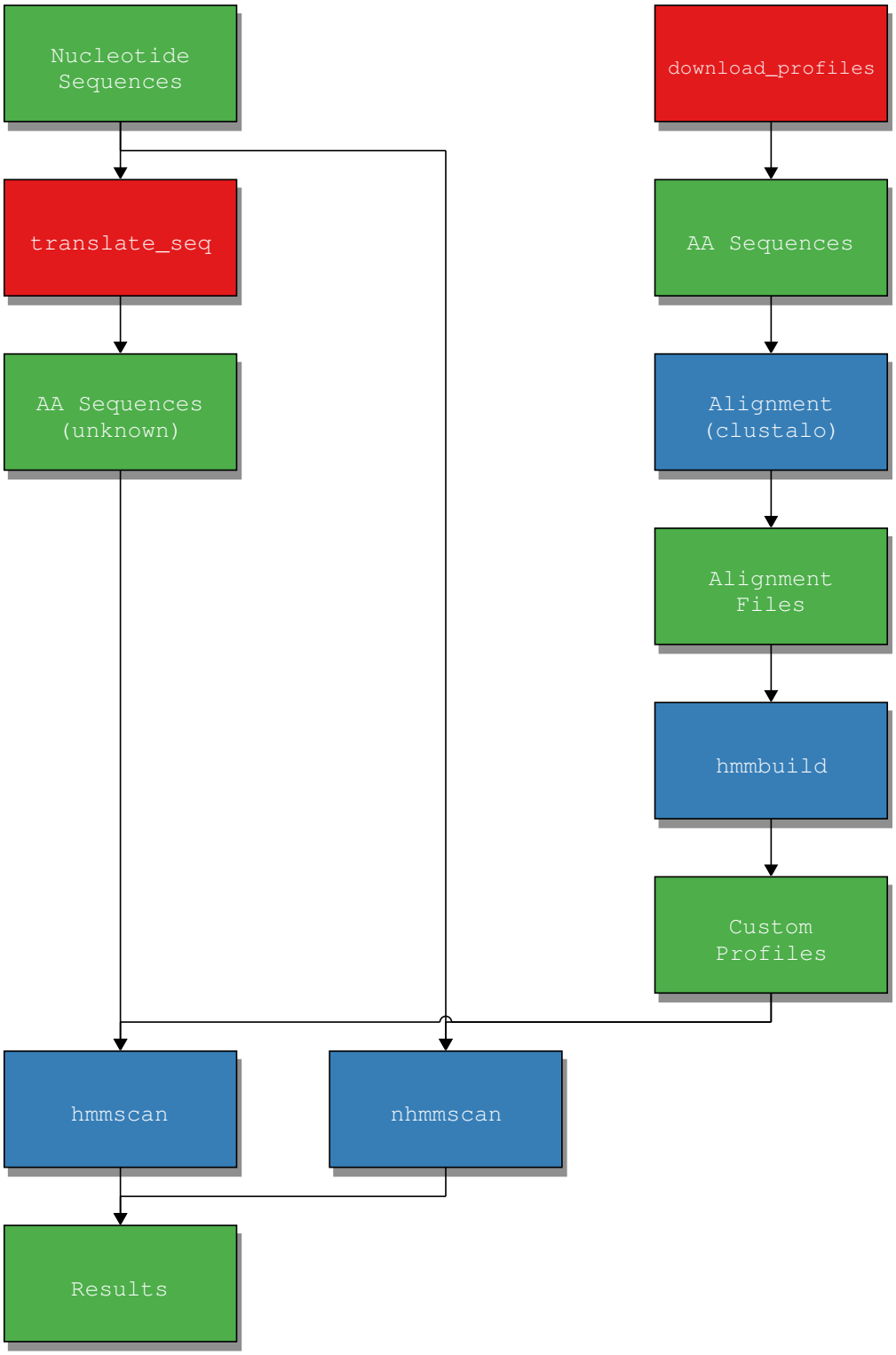
At the moment, the script uses Kegg Orthologs as the ortholog database.

> **Warning:** Some taxa may still black listed, because they are not relevant to the rumen microbiome. If you find such thing to occur to you, please contact me or open an issue on the repository.

4.12.3 Required Data

The script requires data from Kegg Orthologs and Uniprot to be downloaded, before it can be used. The script **download_data** (*download-data - Download Taxonomy from NCBI*) automates the process.

4.12.4 Workflow for Custom Profiles



The process of building the profiles to be used with HMMER is a step that involves several tasks (illustrated in the Workflow above):

1. download of data

2. alignment of sequences
3. conversion in HMMER profiles.

The first step involves, for all ortholog genes, to download all sequences available for each taxon level of interest: this will produce a series of file which contain the amino-acid sequences for each tuple gene-taxon. This sctipt, *download_profiles* can be used. The aminoacidic sequences downloaded are then aligned using Clustal Omega (or other) and for each alignment a profile is built.

HMMER required the use of aminoacidic sequences, to be match against the profiles. The *translate_seq* script can be used to translate nucleotidic sequences into aminoacidic ones. However, the last version of HMMER should be able to match nucleotidic sequences, but it was not tested by us. The example Workflow above illustrate that.

Building profiles in this way, by going through all ortholog genes and choosing the taxon level desired, opens the possibility of incrementally refining the profiling of a metagenome without having to rerun all profiles again, as only the new ones need to be run. Filtering the all the results is a much faster operation.

4.12.5 Usage

The default behaviour is to download all Kegg Orthologs for all taxa in the given taxonomy. Taxa can be filtered by both lineage (e.g. archaea, carnivora, etc.) and rank (e.g. genus, family, etc.). Another option is to specify the KO and taxa IDs to download.

Taxa Filters

The way a taxon is specified is through a few different rules:

- specific **taxon ids** in uniprot
- a specific **taxon rank** (e.g.: genus, phylum, etc.)
- optional lineage filter: the lineage filter make sure that the name specified is included in the lineage attribute in the taxonomy.

As an example, if the rank chosen is genus, and the lineage option is set to archaea, only the taxa whose rank is genus and that belong to the archaea subtree will be downloaded:

```
$ download_profiles -m EMAIL -r genus -l archaea mg_data/kegg.pickle \
-t mg_data/taxonomy.pickle
```

This allows to customise the level of specificity that we want in profiling and make the process of downloading faster. For metagenomic data, a good start is mixing different taxon ranks, using the order or genus for the genes and then specifying a lineage of interest.

Because each profile is indipendent from each other, it's useful to start the download with a certain rank and then run the profiling. During the profiling a new download can be started and so on.

Specific Genes and Taxa

It is possible to download only specific taxa and KO and can be done using the *-i* and *-ko* respectively. When *-ko* is used, loading Kegg Data with *-k* is not required and it is up to the user to ensure the correct genes or taxa.

An example to download only KO from 3 different taxa:

```
$ download_profiles -v -m EMAIL -ko K00016 -i 9611 9645 9682 \
-t mg_data/taxonomy.pickle
```

The same example using taxa filtering, instead (at the time of writing):

```
$ download_profiles -v -m EMAIL -ko K00016 -r genus -l carnivora \
-t mg_data/taxonomy.pickle
```

4.12.6 Changes

Changed in version 0.2.1: added *-ko* option, resolved issues caused by changes in library

4.12.7 Options

Download KO sequences from Uniprot

```
usage: download_profiles [-h] [-o OUTPUT_DIR] [-k KEGG_DATA] -m EMAIL
                         [-t TAXON_DATA] [-r TAXON_RANK] [-l LINEAGE]
                         [-i TAXON_ID [TAXON_ID ...]] [-ko KO_ID [KO_ID ...]]
                         [-R] [-a] [-v | --quiet] [--cite] [--manual]
                         [--version]
```

Named Arguments

-o, --output-dir	directory in which to store the downloaded files
	Default: "profile_files"
-k, --kegg-data	pickle file containing Kegg data
	Default: "data/kegg.pickle"
-m, --email	email address to use for Uniprot communications
-t, --taxon-data	pickle file containing taxonomy data
	Default: "data/taxonomy.pickle"
-r, --taxon-rank	taxon rank to download
-l, --lineage	lineage for filtering (e.g. archaea)
-i, --taxon-id	id(s) of taxa to download. If specified take precedence over lineage-rank
-ko, --ko-id	KO id(s)to download. If specified option -k is not needed
-R, --only-reviewed	Only download reviewed sequences
	Default: False
-a, --all-path	Download all KO from Kegg - exclude blacklist
	Default: False
-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

4.13 sampling-utils - Resampling Utilities

4.13.1 Overview

New in version 0.3.1.

Resampling Utilities

sample command

This command samples from a Fasta or FastQ file, based on a probability defined by the user (0.001 or 1 / 1000 by default, *-r* parameter), for a maximum number of sequences (100,000 by default, *-x* parameter). By default 1 sample is extracted, but as many as desired can be taken, by using the *-n* parameter.

The sequence file in input can be either be passed to the standard input or as last parameter on the command line. By defult a Fasta is expected, unless the *-q* parameter is passed.

The *-p* parameter specifies the prefix to be used, and if the output files can be gzipped using the *-z* parameter.

4.13.2 Options

Sampling utilities

```
usage: sampling-utils [-h] [-v | --quiet] [--cite] [--manual] [--version]
                      {sample} ...
```

Named Arguments

-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

Sub-commands:

sample

Sample a Fasta/FastQ

```
sampling-utils sample [-h] [-p PREFIX] [-n NUMBER] [-r PROB] [-x MAX_SEQ] [-q]
                      [-z] [-v | --quiet] [--cite] [--manual] [--version]
                      [input_file]
```

Positional Arguments

input_file	Input FASTA file, defaults to stdin
	Default: -

Named Arguments

-p, --prefix	Prefix for the file name(s) in output
	Default: "sample"
-n, --number	Number of samples to take
	Default: 1

-r, --prob	Probability of picking a sequence
	Default: 0.001
-x, --max-seq	Maximum number of sequences
	Default: 100000
-q, --fastq	The input file is a fastq file
	Default: False
-z, --gzip	gzip output files
	Default: False
-v, --verbose	more verbose - includes debug messages
	Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

4.14 download-data - Download Taxonomy from NCBI

A bash script called **download-taxonomy.sh** is installed with MGKit. The script downloads the required file (taxdump.tar.gz[63]) form the NCBI ftp taxonomy directory, using **wget**.

Note: If the file is found then it is not downladed. This is handy in case wget is not installed on the system by default (e.g. MacOS X)

The file is then decompressed using **tar** and a *files.txt* file created to be used by a simple script in the same directory. Both the directory and *files.txt* are deleted at the end, but not **taxdump.tar.gz**. A taxonomy file **taxonomy.pickle** is created in the same directory.

4.15 Download Taxa IDs from Uniprot

A script is included to download and prepare a tab separated list of all Taxa IDs associated with Uniprot IDs, so it can be used with *add-gff-info - Add informations to GFF annotations*. The script is called *download-uniprot-taxa.sh* and is installed with MGKit. By default both SwissProt and TrEMBL IDs are downloaded, but passing either *sp* or *SP* will download only SwissProt. The output file is called *uniprot-taxa*

4.16 Download Taxa IDs from NCBI

A script is included to download and prepare a tab separated list of all Taxa IDs associated with NCBI (GenBank) IDs, so it can be used with *add-gff-info - Add informations to GFF annotations*. The script is called *download-ncbi-taxa.sh* and is installed with MGKit. By default *nt* (nucleotide) IDs are downloaded, but passing either *prot* or *PROT* will download *nr* (protein) IDs. The output file is called *ncbi-nucl-taxa.gz* or *ncbi-prot-taxa.gz* depending of the downloaded data.

[63] ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdump.tar.gz

4.17 Download Required Data (Deprecated)

The scripts downloads the data that is used by the framework for some of its functions. It's mostly a shortcut to call the download_data function that is present in every module that is in the package mappings and in the kegg module.

The only option required, is the email contact for the person using the script; this is used to make sure that the API requirements in Uniprot are fullfilled and they can contact the person using the script is any problem arise.

Note: The default behavior is to download first the taxonomy data form Uniprot, Kegg and additional mapping data.

4.17.1 Taxonomy

It is downloaded from Uniprot and build a data structure that is used by several scripts and function in the package. The download can take some time.

> **Warning:** if only the taxonomy is to be downloaded, both the *-x* and *-p* options must be passed to the script.

4.17.2 Kegg Onthologs

The Kegg data is the only "required" data at the moment, because it's used to download the sequence data (via the donwload_profiles script) for the profiling. It is the only data that can't be saved unless it's fully downloaded.

Kegg data is required by the mappers currently supported, and its download takes longer. The mappers handle timeouts and if exceptions are raised the data is saved and the download is resumed when the script is started again.

4.17.3 Other Mappings

The other mappings (from KO) are downloaded by default and this can be excluded by using the *-p* option. Mappings for Gene Onthology, eggNOG and CaZy are downloaded.

As the download of the mappings can take a lot of time, or break because of the number of requests to the web sites, checkpoints are saved often, so it can resumed at a later time.

> **Warning:** the Gene Onthology module has specific requirements, so if they are not downloaded

4.17.4 Options

SNPs analysis, requires a vcf file and SNPDat results

```
usage: download_data [-h] [-o OUTPUT_DIR] [-k KEGG] [-c CAZY] [-g GO]
                     [-e EGGNOG] [-t TAXONOMY] [-p] [-x] -m EMAIL
                     [-v | --quiet] [--cite] [--manual] [--version]
```

Named Arguments

-o, --output-dir Ouput directory

Default: "mg_data"

-k, --kegg	Kegg data file name Default: "kegg.pickle"
-c, --cazy	CaZy data file name Default: "cazy.pickle"
-g, --go	Gene Onthology data file name Default: "go.pickle"
-e, --eggnog	eggNOG data file name Default: "eggnog.pickle"
-t, --taxonomy	Taxonomy data file name Default: "taxonomy.pickle"
-p, --no-mappings	Use to not download Mapping data Default: True
-x, --only-taxonomy	Use to only download Taxonomy data, no Kegg data Default: True
-m, --email	email address to use for Uniprot communications
-v, --verbose	more verbose - includes debug messages Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

4.18 translate_seq - Translate Nucleotides to Aminoacids (Deprecated)

4.18.1 Overview

Deprecated in favour of a similar command in *fasta-utils - Fasta Utilities*

Translate nucleotidic sequence in amino acidic sequences

4.18.2 Options

Translate sequences from nucleotidic to amino acidic

```
usage: translate_seq [-h]
                     [-t {bac_plt,drs_mit,inv_mit,prt_mit,universal,vt_mit,yst_alt,
↪yst_mit}]
                     [-p PROCESSORS] [-n BUFFER_SIZE] [-v | --quiet] [--cite]
                     [--manual] [--version]
                     [input_file] [output_file]
```

Positional Arguments

input_file	input file with aa sequences Default: -
output_file	output file with aa sequences Default: <open file '<stdout>', mode 'w' at 0x7fc92a32b150>

Named Arguments

-t, --trans-table	Possible choices: bac_plt, drs_mit, inv_mit, prt_mit, universal, vt_mit, yst_alt, yst_mit translation table Default: "universal"
-p, --processors	Number of processors to use Default: 1
-n, --buffer-size	Number of sequences to read/write at a time Default: 50000
-v, --verbose	more verbose - includes debug messages Default: 20
--quiet	less verbose - only error and critical messages
--cite	Show citation for the framework
--manual	Show the script manual
--version	show program's version number and exit

CHAPTER 5

Example Notebooks

Example notebooks are included about using the library

CHAPTER 6

MGKit GFF Specifications

The GFF produced with MGKit follows the conventions of GFF/GTF files but it provides some additional fields in the 9th column which translate to a Python dictionary when an annotation is loaded into an `Annotation` instance.

The 9th column is a list of **key=value** item, separated by a semicolon (;); each value is also expected to be quoted with double quotes and the values to not include a semicolon or other characters that can make the parsing difficult. MGKit uses `urllib.quote()` to encode those characters and also " ()/". The *`mgkit.io.gff.from_gff()`* uses `urllib.unquote()` to set the values.

Warning: As the last column translates to a dictionary in the data structures, duplicate keys are not allowed. *`mgkit.io.gff.from_gff()`* raises an exception if any are found.

6.1 Reserved Values

Any key can be added to a GFF annotation, but MGKit expects a few key to be in the GFF annotation as summarised in the following tables.

Table 1: Reserved values, used by the scripts

Key	Value	Explanation
gene_id	any string	used to identify the gene predicted
db	any string, like UNIPROT-SP, UNIPROT-TR, NCBI-NT	identifies the database used to make the gene_id prediction
taxon_db	any string, like UNIPROT-SP, UNIPROT-TR, NCBI-NT	identifies the database used to make the taxon_id prediction
dbq	integer	identifies the quality of the database, used when filtering annotations
taxon_id	integer	identifies the annotation taxon, NCBI taxonomy is used
uid	string	unique identifier for the annotation, any string is accepted but a value is assigned by using `uuid.uuid4()`[64]
cov and {any}_cov	integer	coverage for the annotation over all samples, keys ending with *_cov* indicates coverage for each sample
exp_syn, exp_nonsyn	integer	used for expected number of synonymous and non-synonymous changes for the annotation

The following keys are added by different scripts and may be used in different scripts or annotation methods.

Table 2: Interpreted Values

Key	Value	Explanation	Used
taxon_name	string	name of the taxon	not used
lineage	string	taxon lineage	not used
EC	comma separated values	list of EC numbers associated to the annotation	used by *mgkit.io.gff.Annotation.get_ec()*
map_{any}	comma separated values	list of mapping to a specific db (e.g. eggNOG -> map_EGGNOG)	used by *mgkit.io.gff.Annotation.get_mapping()*
counts_{any}	float	Stores the count data for a sample (e.g. counts_Sample1)	used by script *add-gff-info*
fpkms_{any}	float	Stores the count data for a sample (e.g. fpkms_Sample1)	used by script *add-gff-info*

[64] https://docs.python.org/3/library/uuid.html#uuid.uuid4

CHAPTER 7

Library Reference

7.1 mgkit package

7.1.1 Subpackages

mgkit.counts package

Submodules

mgkit.counts.func module

New in version 0.1.13.

Misc functions for count data

mgkit.counts.func.**batch_load_htseq_counts**(*count_files*, *samples=None*, *cut_name=None*)

Loads a list of htseq count result files and returns a DataFrame (IDxSAMPLE)

The sample names are names are the file names if *samples* and *cut_name* are *None*, supplying a list of sample names with *samples* is the preferred way, and *cut_name* is used for backward compatibility and as an option in cases a string replace is enough.

Parameters

- **count_files** (*file or str*[65]) – file handle or string with file name
- **samples** (*iterable*) – list of sample names, in the same order as *count_files*
- **cut_name** (*str*[66]) – string to delete from the the file names to get the sample names

Returns with sample names as columns and gene_ids as index

Return type pandas.DataFrame

mgkit.counts.func.**filter_counts**(*counts_iter*, *info_func*, *gfilters=None*, *tfilters=None*)

Returns counts that pass filters for each *uid* associated *gene_id* and *taxon_id*.

Parameters

[65] https://docs.python.org/3/library/stdtypes.html#str
[66] https://docs.python.org/3/library/stdtypes.html#str

- **counts_iter** (*iterable*) – iterator that yields a tuple (uid, count)
- **info_func** (*func*) – function accepting a *uid* that returns a tuple *(gene_id, taxon_id)*
- **gfilters** (*iterable*) – list of filters to apply to each *uid* associated *gene_id*
- **tfilters** (*iterable*) – list of filters to apply to each *uid* associated *taxon_id*

Yields *tuple – (uid, count)* that pass filters

mgkit.counts.func.**from_gff**(*annotations*, *samples*, *ann_func=None*, *sample_func=None*)

New in version 0.3.1.

Loads count data from a GFF file, only for the requested samples. By default the function returns a DataFrame where the index is the *uid* of each annotation and the columns the requested samples.

This can be customised by supplying *ann_func* and *sample_func*. *sample_func* is a function that accept a sample name and is expected to return a string or a tuple. This will be used to change the columns in the DataFrame. *ann_func* must accept an *mgkit.io.gff.Annotation* instance and return an iterable, with each iteration yielding either a single element or a tuple (for a MultiIndex DataFrame), each element yielded will have the count of that annotation added to.

Parameters

- **annotation** (*iterable*) – iterable yielding annotations
- **samples** (*iterable*) – list of samples to keep
- **ann_func** (*func*) – function used to customise the output
- **sample_func** (*func*) – function to customise the column elements

Returns dataframe with the count data, columns are the samples and rows the annotation counts (unless mapped with *ann_func*)

Return type DataFrame

Exmples: Assuming we have a list of *annotations* and sample SAMPLE1 and SAMPLE2 we can obtain the count table for all annotations with this

```
>>> from_gff(annotations, ['SAMPLE1', 'SAMPLE2'])
```

Assuming we want to group the samples, for example treatment1, treatment2 and control1, control2 into a MultiIndex DataFrame column

```
>>> sample_func = lambda x: ('T' if x.startswith('t') else 'C', x)
>>> from_gff(annotations, ['treatment1', 'treatment2', 'control1',
↪'control2'], sample_func=sample_func)
```

Annotations can be mapped to other levels for example instead of using the *uid* that is the default, it can be mapped to the gene_id, taxon_id information that is included in the annotation, resulting in a MultiIndex index for the rows, with (gene_id, taxon_id) as key.

```
>>> ann_func = lambda x: [(x.gene_id, x.taxon_id)]
>>> from_gff(annotations, ['SAMPLE1', 'SAMPLE2'], ann_func=ann_func)
```

mgkit.counts.func.**get_uid_info**(*info_dict*, *uid*)

Simple function to get a value from a dictionary of tuples (gene_id, taxon_id)

mgkit.counts.func.**get_uid_info_ann**(*annotations*, *uid*)

Simple function to get a value from a dictionary of annotations

mgkit.counts.func.**load_counts_from_gff**(*annotations*, *elem_func=<function <lambda>>*, *sample_func=None*, *nozero=True*)

New in version 0.2.5.

Loads counts for each annotations that are stored into the annotation *counts_* attributes. Annotations with a total of 0 counts are skipped by default (nozero=True), the row index is set to the *uid* of the annotation and the column to the sample name. The functions used to transform the indices expect the annotation (for the row, *elem_func*) and the sample name (for the column, *sample_func*).

Parameters

- **annotations** (*iter*) – iterable of annotations
- **elem_func** (*func*) – function that accepts an annotation and return a str/int for a Index or a tuple for a MultiIndex, defaults to returning the *uid* of the annotation
- **sample_func** (*func, None*) – function that accepts the sample name and returns tuple for a MultiIndex. Defaults to *None* so no transformation is performed
- **nozero** (*bool*[67]) – if *True*, annotations with no counts are skipped

mgkit.counts.func.**load_deseq2_results**(*file_name*, *taxon_id=None*)

New in version 0.1.14.

Reads a CSV file output with DESeq2 results, adding a taxon_id to the index for concatenating multiple results from different taxonomic groups.

Parameters **file_name** (*str*[68]) – file name of the CSV

Returns a MultiIndex DataFrame with the results

Return type pandas.DataFrame

mgkit.counts.func.**load_htseq_counts**(*file_handle*, *conv_func=<type 'int'>*)

Changed in version 0.1.15: added *conv_func* parameter

Loads an HTSeq-count result file

Parameters

- **file_handle** (*file or str*[69]) – file handle or string with file name
- **conv_func** (*func*) – function to convert the number from string, defaults to *int*, but *float* can be used as well

Yields *tuple* – first element is the gene_id and the second is the count

mgkit.counts.func.**load_sample_counts**(*info_dict*, *counts_iter*, *taxonomy*, *inc_anc=None*, *rank=None*, *gene_map=None*, *ex_anc=None*, *include_higher=True*, *cached=True*, *uid_used=None*)

Changed in version 0.1.14: added *cached* argument

Changed in version 0.1.15: added *uid_used* parameter

Changed in version 0.2.0: info_dict can be a function

Reads sample counts, filtering and mapping them if requested. It's an example of the usage of the above functions.

Parameters

- **info_dict** (*dict*[70]) – dictionary that has *uid* as key and *(gene_id, taxon_id)* as value. In alternative a function that accepts a *uid* as sole argument and returns *(gene_id, taxon_id)*
- **counts_iter** (*iterable*) – iterable that yields a *(uid, count)*
- **taxonomy** – taxonomy instance

[67] https://docs.python.org/3/library/functions.html#bool
[68] https://docs.python.org/3/library/stdtypes.html#str
[69] https://docs.python.org/3/library/stdtypes.html#str
[70] https://docs.python.org/3/library/stdtypes.html#dict

- **inc_anc** (*int*[71], *list*[72]) – ancestor taxa to include
- **rank** (*str*[73]) – rank to which map the counts
- **gene_map** (*dict*[74]) – dictionary with the gene mappings
- **ex_anc** (*int*[75], *list*[76]) – ancestor taxa to exclude
- **include_higher** (*bool*[77]) – if False, any rank different than the requested one is discarded
- **cached** (*bool*[78]) – if *True*, the function will use *mgkit.simple_cache.memoize* to cache some of the functions used
- **uid_used** (*None*, *dict*[79]) – an empty dictionary in which to store the *uid* that were assigned to each key of the returned pandas.Series. If *None*, no information is saved

Returns array with MultiIndex *(gene_id, taxon_id)* with the filtered and mapped counts

Return type pandas.Series

mgkit.counts.func.**load_sample_counts_to_genes**(*info_func*, *counts_iter*, *taxonomy*, *inc_anc=None*, *gene_map=None*, *ex_anc=None*, *cached=True*, *uid_used=None*)

New in version 0.1.14.

Changed in version 0.1.15: added *uid_used* parameter

Reads sample counts, filtering and mapping them if requested. It's a variation of *load_sample_counts()*, with the counts being mapped only to each specific gene_id. Another difference is the absence of any assumption on the first parameter. It is expected to return a (gene_id, taxon_id) tuple.

Parameters

- **info_func** (*callable*) – any callable that accept an *uid* as the only parameter and and returns *(gene_id, taxon_id)* as value
- **counts_iter** (*iterable*) – iterable that yields a *(uid, count)*
- **taxonomy** – taxonomy instance
- **inc_anc** (*int*[80], *list*[81]) – ancestor taxa to include
- **rank** (*str*[82]) – rank to which map the counts
- **gene_map** (*dict*[83]) – dictionary with the gene mappings
- **ex_anc** (*int*[84], *list*[85]) – ancestor taxa to exclude
- **cached** (*bool*[86]) – if *True*, the function will use *mgkit.simple_cache.memoize* to cache some of the functions used

[71] https://docs.python.org/3/library/functions.html#int
[72] https://docs.python.org/3/library/stdtypes.html#list
[73] https://docs.python.org/3/library/stdtypes.html#str
[74] https://docs.python.org/3/library/stdtypes.html#dict
[75] https://docs.python.org/3/library/functions.html#int
[76] https://docs.python.org/3/library/stdtypes.html#list
[77] https://docs.python.org/3/library/functions.html#bool
[78] https://docs.python.org/3/library/functions.html#bool
[79] https://docs.python.org/3/library/stdtypes.html#dict
[80] https://docs.python.org/3/library/functions.html#int
[81] https://docs.python.org/3/library/stdtypes.html#list
[82] https://docs.python.org/3/library/stdtypes.html#str
[83] https://docs.python.org/3/library/stdtypes.html#dict
[84] https://docs.python.org/3/library/functions.html#int
[85] https://docs.python.org/3/library/stdtypes.html#list
[86] https://docs.python.org/3/library/functions.html#bool

- **uid_used** (*None*, *dict*[87]) – an empty dictionary in which to store the *uid* that were assigned to each key of the returned pandas.Series. If *None*, no information is saved

Returns array with Index *gene_id* with the filtered and mapped counts

Return type pandas.Series

mgkit.counts.func.**load_sample_counts_to_taxon**(*info_func*, *counts_iter*, *taxonomy*, *inc_anc=None*, *rank=None*, *ex_anc=None*, *include_higher=True*, *cached=True*, *uid_used=None*)

New in version 0.1.14.

Changed in version 0.1.15: added *uid_used* parameter

Reads sample counts, filtering and mapping them if requested. It's a variation of *load_sample_counts()*, with the counts being mapped only to each specific taxon. Another difference is the absence of any assumption on the first parameter. It is expected to return a (gene_id, taxon_id) tuple.

Parameters

- **info_func** (*callable*) – any callable that accept an *uid* as the only parameter and and returns *(gene_id, taxon_id)* as value
- **counts_iter** (*iterable*) – iterable that yields a *(uid, count)*
- **taxonomy** – taxonomy instance
- **inc_anc** (*int*[88], *list*[89]) – ancestor taxa to include
- **rank** (*str*[90]) – rank to which map the counts
- **ex_anc** (*int*[91], *list*[92]) – ancestor taxa to exclude
- **include_higher** (*bool*[93]) – if False, any rank different than the requested one is discarded
- **cached** (*bool*[94]) – if *True*, the function will use *mgkit.simple_cache.memoize* to cache some of the functions used
- **uid_used** (*None*, *dict*[95]) – an empty dictionary in which to store the *uid* that were assigned to each key of the returned pandas.Series. If *None*, no information is saved

Returns array with Index *taxon_id* with the filtered and mapped counts

Return type pandas.Series

mgkit.counts.func.**map_counts**(*counts_iter*, *info_func*, *gmapper=None*, *tmapper=None*, *index=None*, *uid_used=None*)

Changed in version 0.1.14: added *index* parameter

Changed in version 0.1.15: added *uid_used* parameter

Maps counts according to the gmapper and tmapper functions. Each mapped gene ID count is the sum of all uid that have the same ID(s). The same is true for the taxa.

Parameters

- **counts_iter** (*iterable*) – iterator that yields a tuple (uid, count)
- **info_func** (*func*) – function accepting a *uid* that returns a tuple *(gene_id, taxon_id)*

[87] https://docs.python.org/3/library/stdtypes.html#dict
[88] https://docs.python.org/3/library/functions.html#int
[89] https://docs.python.org/3/library/stdtypes.html#list
[90] https://docs.python.org/3/library/stdtypes.html#str
[91] https://docs.python.org/3/library/functions.html#int
[92] https://docs.python.org/3/library/stdtypes.html#list
[93] https://docs.python.org/3/library/functions.html#bool
[94] https://docs.python.org/3/library/functions.html#bool
[95] https://docs.python.org/3/library/stdtypes.html#dict

- **gmapper** (*func*) – fucntion that accepts a *gene_id* and returns a list of mapped IDs
- **tmapper** (*func*) – fucntion that accepts a *taxon_id* and returns a new *taxon_id*
- **index** (*None, str*[96]) – if None, the index of the Series if *(gene_id, taxon_id)*, if a str, it can be either *gene* or *taxon*, to specify a single value
- **uid_used** (*None, dict*[97]) – an empty dictionary in which to store the *uid* that were assigned to each key of the returned pandas.Series. If *None*, no information is saved

Returns array with MultiIndex *(gene_id, taxon_id)* with the mapped counts

Return type pandas.Series

mgkit.counts.func.**map_counts_to_category**(*counts*, *gene_map*, *nomap=False*, *nomap_id='NOMAP'*)

Used to map the counts from a certain gene identifier to another. Genes with no mappings are not counted, unless *nomap=True*, in which case they are counted as *nomap_id*.

Parameters

- **counts** (*iterator*) – an iterator that yield a tuple, with the first value being the gene_id and the second value the count for it
- **gene_map** (*dictionary*) – a dictionary whose keys are the gene_id yield by *counts* and the values are iterable of mapping identifiers
- **nomap** (*bool*[98]) – if False, counts for genes with no mappings in *gene_map* are discarded, if True, they a counted as *nomap_id*
- **nomap_id** (*str*[99]) – name of the mapping for genes with no mappings

Returns mapped counts

Return type pandas.Series

mgkit.counts.func.**map_gene_id_to_map**(*gene_map*, *gene_id*)

Function that extract a list of gene mappings from a dictionary and returns an empty list if the *gene_id* is not found.

mgkit.counts.func.**map_taxon_id_to_rank**(*taxonomy*, *rank*, *taxon_id*, *include_higher=True*)

Maps a *taxon_id* to the request taxon rank. Returns *None* if *include_higher* is False and the found rank is not the one requested.

Internally uses *mgkit.taxon.UniprotTaxonomy.get_ranked_taxon()*

Parameters

- **taxonomy** – taxonomy instance
- **rank** (*str*[100]) – taxonomic rank requested
- **taxon_id** (*int*[101]) – taxon_id to map
- **include_higher** (*bool*[102]) – if False, any rank different than the requested one is discarded

Returns if the mapping is successful, the ranked taxon_id is returned, otherwise *None* is returned

Return type (int[103], None)

[96] https://docs.python.org/3/library/stdtypes.html#str
[97] https://docs.python.org/3/library/stdtypes.html#dict
[98] https://docs.python.org/3/library/functions.html#bool
[99] https://docs.python.org/3/library/stdtypes.html#str
[100] https://docs.python.org/3/library/stdtypes.html#str
[101] https://docs.python.org/3/library/functions.html#int
[102] https://docs.python.org/3/library/functions.html#bool
[103] https://docs.python.org/3/library/functions.html#int

mgkit.counts.scaling module

Scaling functions for counts

mgkit.counts.scaling.**scale_deseq**(*dataframe*)
New in version 0.1.13.

Scale a dataframe using the deseq scaling. Uses *scale_factor_deseq()*

mgkit.counts.scaling.**scale_factor_deseq**(*dataframe*)
New in version 0.1.13.

Returns the scale factor according to he deseq paper. The columns of the dataframe are the samples.

size factor $\hat{s}_j$ for sample *j* (from DESeq paper).

$$\hat{s}_j = median_i(\frac{k_{ij}}{(\prod_{v=1}^{m} k_{iv})^{1/m}})$$

mgkit.counts.scaling.**scale_rpkm**(*dataframe*, *gene_len*)
New in version 0.1.14.

Perform an RPKM scaling of the pandas dataframe/series supplied using the *gene_len* series containing the gene sizes for all elements of *dataframe*

$$RPKM = \frac{10^9 \cdot C}{N \cdot L}$$

Module contents

mgkit.db package

Submodules

mgkit.db.dbm module

mgkit.db.mongo module

Module contents

mgkit.filter package

Submodules

mgkit.filter.common module

Common consts/data for package filter

exception mgkit.filter.common.**FilterFails**
Bases: exceptions.Exception

Raised if a filter fails

mgkit.filter.gff module

GFF filtering

mgkit.filter.gff.**choose_annotation**(*ann1*, *ann2*, *overlap=100*, *choose_func=None*)

New in version 0.1.12.

Given two *mgkit.io.gff.Annotation*, if one of of the two annotations either is contained in the other or they overlap for at least a *overlap* number of bases, *choose_func* will be applied to both. The result of *choose_func* is the the annotation to be discarderd. It returns *None* if the annotations should be both kept.

No checks are made to ensure that the two annotations are on the same sequence and strand, as the *intersect* method of *mgkit.io.gff.Annotation* takes care of them.

Parameters

- **ann1** – instance of *mgkit.io.gff.Annotation*
- **ann2** – instance of *mgkit.io.gff.Annotation*
- **overlap** (*int*[104], *float*[105]) – number of bases overlap that trigger the filtering
- **choose_func** (*None*, *func*) – function that accepts *ann1* and *ann2* and return the one to be discarded or None if both are accepted

Returns returns either the *mgkit.io.gff.Annotation* to be discarded or None, which is the result of *choose_func*

Return type (None, *Annotation*)

Note: If *choose_func* is *None*, the default function is used:

```
lambda a1, a2: min(a1, a2, key=lambda el: (el.dbq, el.bitscore,
                    len(el)))
```

In order of importance the db quality, the bitscore and the length. The annotation with the lowest tuple value is the one to discard.

---

mgkit.filter.gff.**filter_annotations**(*annotations*, *choose_func=None*, *sort_func=None*, *reverse=True*)

New in version 0.1.12.

Filter an iterable of *mgkit.io.gff.Annotation* instances sorted using *sort_func* as key in *sorted* and if the order is to be *reverse*; it then applies *choose_func* on all possible pair combinations, using itertools.combinations.

By default *choose_func* is *choose_annotation()* with the default values, the list of annotation is sorted by bitscore, from the highest to the lowest value.

Parameters

- **annotations** (*iterable*) – iterable of *mgkit.io.gff.Annotation* instances
- **choose_func** (*func*, *None*) – function used to select the *losing* annotation; if *None*, it will be *choose_annotation()* with default values
- **sort_func** (*func*, *None*) – by default the sorting key is the bitscore of the annotations
- **reverse** (*bool*[106]) – passed to *sorted*, by default is reversed

Returns a set with the annotations that pass the filtering

Return type set[107]

[104] https://docs.python.org/3/library/functions.html#int
[105] https://docs.python.org/3/library/functions.html#float
[106] https://docs.python.org/3/library/functions.html#bool
[107] https://docs.python.org/3/library/stdtypes.html#set

mgkit.filter.gff.**filter_attr_num**(*annotation*, *attr=None*, *value=None*, *greater=True*)

Checks if an annotation *attr* dictionary contains a key whose value is greater than or equal, or lower than or equal, for the requested value

Parameters

- **annotation** – *mgkit.io.gff.Annotation* instance
- **attr** (*str*[108]) – key in the *mgkit.io.gff.Annotation.attr* dictionary
- **value** (*int*[109]) – the value to which we need to compare
- **greater** (*bool*[110]) – if True the value must be equal or greater than and if False equal of lower than

Returns True if the test passes

Return type bool[111]

mgkit.filter.gff.**filter_attr_num_s**(*annotation*, *attr=None*, *value=None*, *greater=True*)

New in version 0.3.1.

Checks if an annotation *attr* dictionary contains a key whose value is greater or lower than the requested value

Parameters

- **annotation** – *mgkit.io.gff.Annotation* instance
- **attr** (*str*[112]) – key in the *mgkit.io.gff.Annotation.attr* dictionary
- **value** (*int*[113]) – the value to which we need to compare
- **greater** (*bool*[114]) – if True the value must be greater than and if False lower than

Returns True if the test passes

Return type bool[115]

mgkit.filter.gff.**filter_attr_str**(*annotation*, *attr=None*, *value=None*, *equal=True*)

Checks if an annotation *attr* dictionary contains a key shose value is equal to, or contains the requested value

Parameters

- **annotation** – *mgkit.io.gff.Annotation* instance
- **attr** (*str*[116]) – key in the *mgkit.io.gff.Annotation.attr* dictionary
- **value** (*int*[117]) – the value to which we need to compare
- **equal** (*bool*[118]) – if True the value must be equal and if False equal value must be contained

Returns True if the test passes

Return type bool[119]

mgkit.filter.gff.**filter_base**(*annotation*, *attr=None*, *value=None*)

Checks if an annotation attribute is equal to the requested value

[108] https://docs.python.org/3/library/stdtypes.html#str
[109] https://docs.python.org/3/library/functions.html#int
[110] https://docs.python.org/3/library/functions.html#bool
[111] https://docs.python.org/3/library/functions.html#bool
[112] https://docs.python.org/3/library/stdtypes.html#str
[113] https://docs.python.org/3/library/functions.html#int
[114] https://docs.python.org/3/library/functions.html#bool
[115] https://docs.python.org/3/library/functions.html#bool
[116] https://docs.python.org/3/library/stdtypes.html#str
[117] https://docs.python.org/3/library/functions.html#int
[118] https://docs.python.org/3/library/functions.html#bool
[119] https://docs.python.org/3/library/functions.html#bool

Parameters

- **annotation** – *mgkit.io.gff.Annotation* instance
- **attr** (*str*[120]) – attribute of the annotation
- **value** – the value that the attribute should be equal to

Returns True if the supplied value is equal to the attribute ot False otherwise

Return type bool[121]

mgkit.filter.gff.**filter_base_num**(*annotation*, *attr=None*, *value=None*, *greater=True*)
Checks if an annotation attribute is greater, equal of lower than the requested value

Parameters

- **annotation** – *mgkit.io.gff.Annotation* instance
- **attr** (*str*[122]) – attribute of the annotation
- **value** (*int*[123]) – the value to which the attribute should be compared to
- **greater** (*bool*[124]) – if True the attribute value must be equal or greater than and if False equal of lower than

Returns True if the test passes

Return type bool[125]

mgkit.filter.gff.**filter_len**(*annotation*, *value=None*, *greater=True*)
Checks if an annotation length is longer, equal of shorter than the requested value

Parameters

- **annotation** – *mgkit.io.gff.Annotation* instance
- **value** (*int*[126]) – the length to which the attribute should be compared to
- **greater** (*bool*[127]) – if True the annotation length must be equal or greater than and if False equal of lower than

Returns True if the test passes

Return type bool[128]

mgkit.filter.gff_old module

Old GFF filtering module

mgkit.filter.gff_old.**choose_annotation**(*ann1*, *ann2*, *threshold*, *only_same_gene=False*, *only_same_strand=True*, *score_func=<function choose_by_score>*)
Choos one annotation, based on the score

mgkit.filter.gff_old.**choose_by_score**(*ann1*, *ann2*)
the winner is the first element of the list

mgkit.filter.gff_old.**filter_by_bit_score**(*threshold*, *annotation*)
Filter based on the bit score of the annotation

[120] https://docs.python.org/3/library/stdtypes.html#str
[121] https://docs.python.org/3/library/functions.html#bool
[122] https://docs.python.org/3/library/stdtypes.html#str
[123] https://docs.python.org/3/library/functions.html#int
[124] https://docs.python.org/3/library/functions.html#bool
[125] https://docs.python.org/3/library/functions.html#bool
[126] https://docs.python.org/3/library/functions.html#int
[127] https://docs.python.org/3/library/functions.html#bool
[128] https://docs.python.org/3/library/functions.html#bool

mgkit.filter.gff_old.**filter_by_description**(*description*, *annotation*)
Filter based on the description of the annotation

mgkit.filter.gff_old.**filter_by_hit_length**(*ko_len*, *perc*, *annotation*)
Filter based on the its length and the profile length of the annotation

mgkit.filter.gff_old.**filter_by_ko_id**(*ko_ids*, *annotation*)
Filter based on the KO id of the annotation

mgkit.filter.gff_old.**filter_by_ko_idx**(*ko_ids*, *annotation*)
Filter based on the KO id (indexed) of the annotation

mgkit.filter.gff_old.**filter_by_reviewed**(*annotation*)
Filter based on the reviewed attribute of the annotation

mgkit.filter.gff_old.**filter_by_score**(*threshold*, *annotation*)
Filter based on the score of the annotation

mgkit.filter.gff_old.**filter_by_seq_id**(*seq_ids*, *annotation*)
Filter based on the sequence containing the annotation

mgkit.filter.gff_old.**filter_by_strand**(*strand*, *annotation*)
Filter based on the strand of the annotation

mgkit.filter.gff_old.**filter_by_taxon**(*taxa*, *annotation*)
Filter based on the taxon name of the annotation

mgkit.filter.gff_old.**filter_overlapping**(*annotations*, *threshold*, *same_gene*, *both_strands*, *choose_func=<function choose_annotation>*, *score_func=<function choose_by_score>*)
Filters annotations checking how much they overlap, in which case the ones with better score are used.

Parameters

- **annotations** (*iterable*) – iterable of GFF annotations
- **threshold** (*float*[129]) – maximum overlap allowed
- **same_gene** (*bool*[130]) – if True only filter genes that have the same id
- **both_strands** (*bool*[131]) – if True checks annotations on both strands
- **choose_func** (*func*) – function used to choose the annotations. Defaults to *choose_annotation()*
- **score_func** (*func*) – function used to decide which annotation to keep if two overlap

Return set annotations that passed the filtering

mgkit.filter.lists module

Module used to filter lists

mgkit.filter.lists.**aggr_filtered_list**(*val_list*, *aggr_func=<function mean>*, *filt_func=<function <lambda>>*)
Aggregate a list of values using 'aggr_func' on a list that passed the filtering in 'filt_func'.

'filt_func' is a function that returns True or False for each value in val_list. If the return value is True, the element is included in the values passed to 'aggr_func'. Internally a list comprehension is used and the result passed to 'aggr_func'

Parameters

[129] https://docs.python.org/3/library/functions.html#float
[130] https://docs.python.org/3/library/functions.html#bool
[131] https://docs.python.org/3/library/functions.html#bool

- **val_list** (*iterable*) – list of values
- **aggr_func** (*func*) – function used to aggregate the list values
- **filt_func** (*func*) – function the return True or False

Returns the result of the applied 'aggr_func'

mgkit.filter.reads module

Some test functions to filter sequences

mgkit.filter.reads.**expected_error_rate**(*qualities*)
Calculate the expected error rate for an array of qualities (converted to probabilities).

mgkit.filter.reads.**trim_by_ee**(*qualities*, *min_length=50*, *threshold=0.5*, *chars=True*, *base=33*)
Trim a sequence based on the expected error rate.

mgkit.filter.taxon module

New in version 0.1.9.

Taxa filtering functions

mgkit.filter.taxon.**filter_by_ancestor**(*taxon_id*, *filter_list=None*, *exclude=False*, *taxonomy=None*)
New in version 0.1.13.

Convenience function for *filter_taxon_by_id_list()*, as explained in the latter example.

mgkit.filter.taxon.**filter_taxon_by_id_list**(*taxon_id*, *filter_list=None*, *exclude=False*, *func=None*)
Filter a taxon_id against a list of taxon ids. Returns True if the conditions of the filter are met.

If func is not None, a function that accepts two values is expected, it should be either a partial *is_ancestor* which only accepts taxon_id and anc_id or another function that behaves the same way.

Note: if func is None, a simple lambda is used to test identity:

```
func = lambda t_id, a_id: t_id == a_id
```

Parameters

- **taxon_id** (*int*[132]) – the taxon id to filter
- **filter_list** (*iterable*) – an iterable with taxon ids
- **exclude** (*bool*[133]) – if the filter is reversed (i.e. included if NOT found)
- **func** (*func or None*) – a function that accepts taxon_id and an anc_id and returns a bool to indicated if anc_id is ancestor of taxon_id. Equivalent to *is_ancestor()*.

Returns

True if the taxon_id is in the filter list (or a descendant of it) False if it's not found. Exclude equal to True reverse the result.

[132] https://docs.python.org/3/library/functions.html#int
[133] https://docs.python.org/3/library/functions.html#bool

Found	Exclude	Return Value
Yes	False	True
No	False	False
Yes	True	False
No	True	True

Return type bool[134]

Example

If using func and assuming that *taxonomy* is an instance of `UniprotTaxonomy` with data loaded:

```
>>> import functools
>>> import mgkit.taxon
>>> func = functools.partial(mgkit.taxon.is_ancestor, taxonomy)
>>> filter_taxon_by_id_list(1200582, [838], func=func)
True
```

Module contents

Package used to store filter functions (unless specific to a package)

mgkit.io package

Submodules

mgkit.io.blast module

Blast routines and parsers

`mgkit.io.blast.`**`add_blast_result_to_annotation`**(*annotation*, *gi_taxa_dict*, *taxonomy*, *threshold=60*)

Adds blast information to a GFF annotation.

Parameters

- **`annotation`** – GFF annotation object
- **`gi_taxa_dict`** (*dict*[135]) – dictionary returned by `parse_gi_taxa_table()`.
- **`taxonomy`** – Uniprot taxonomy, used to add the taxon name to the annotation

`mgkit.io.blast.`**`parse_accession_taxa_table`**(*file_handle*, *acc_ids=None*, *key=1*, *value=2*, *num_lines=1000000*, *no_zero=True*)

New in version 0.2.5.

Changed in version 0.3.0: added *no_zero*

This function superseeds `parse_gi_taxa_table()`, since NCBI is deprecating the GIDs in favor of accessions like *X53318*. The new file can be found at the NCBI *ftp://ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid*, for DNA sequences (*nt* DB) *nucl_gb.accession2taxid.gz*.

The file contains 4 columns, the first one is the accession without its version, the second one includes the version, the third column is the taxonomic identifier and the fourth is either the old GID or **na**.

[134] https://docs.python.org/3/library/functions.html#bool
[135] https://docs.python.org/3/library/stdtypes.html#dict

The column used as key is the *second*, since by default the fasta headers used in NCBI DBs use the versioned identifier. To use the GID as key, the *key* parameter can be set to 3, but if no identifier is found (*na* as per the file README), the line is skipped.

Parameters

- **file_handle** (*str*[136], *file*) – file name or open file handle
- **acc_ids** (*None*, *list*[137]) – if it's not *None* only the keys included in the passed *acc_ids* list will be returned
- **key** (*int*[138]) – 0-based index for the column to use as accession. Defaults to the versioned accession that is used in GenBank fasta files.
- **num_lines** (*None*, *int*[139]) – number of which a message is logged. If None, no message is logged
- **no_zero** (*bool*[140]) – if True (default) a key with taxon_id of 0 is not yield

---

Note: GIDs are being phased out in September 2016: http://www.ncbi.nlm.nih.gov/news/03-02-2016-phase-out-of-GI-numbers/

---

mgkit.io.blast.**parse_blast_tab**(*file_handle*, *seq_id=0*, *ret_col=(0, 1, 2, 6, 7, 11)*, *key_func=None*, *value_funcs=None*)

New in version 0.1.12.

Parses blast output tab format, returning for each line a key (the query id) and the columns requested in a tuple.

Parameters

- **file_handle** (*file*) – file name or file handle for the blast ouput
- **seq_id** (*int*[141]) – index for the column which has the query id
- **ret_col** (*list*[142], *None*) – list of indexes for the columns to be returned or *None* if all columns must be returned
- **key_func** (*None*, *func*) – function to transform the query id value in the key returned. If *None*, the query id is used
- **value_funcs** (*None*, *list*[143]) – list of functions to transform the value of all the requested columns. If *None* the values are not converted

Yields *tuple* – iterator of tuples with the first element being the query id after key_func is applied, if requested and the second element of the tuple is a tuple with the requested columns *ret_col*

[136] https://docs.python.org/3/library/stdtypes.html#str
[137] https://docs.python.org/3/library/stdtypes.html#list
[138] https://docs.python.org/3/library/functions.html#int
[139] https://docs.python.org/3/library/functions.html#int
[140] https://docs.python.org/3/library/functions.html#bool
[141] https://docs.python.org/3/library/functions.html#int
[142] https://docs.python.org/3/library/stdtypes.html#list
[143] https://docs.python.org/3/library/stdtypes.html#list

Table 1: BLAST+ used with *-outfmt 6*, default columns

column index	description
0	query name
1	subject name
2	percent identities
3	aligned length
4	number of mismatched positions
5	number of gap positions
6	query sequence start
7	query sequence end
8	subject sequence start
9	subject sequence end
10	e-value
11	bit score

mgkit.io.blast.**parse_fragment_blast**(*file_handle*, *bitscore=40.0*)

New in version 0.1.13.

Parse the output of a BLAST output where the sequences are the single annotations, so the sequence names are the *uid* of the annotations.

The only returned values are the best hits, maxed by bitscore and identity.

Parameters

- **file_handle** (*str*[144], *file*) – file name or open file handle
- **bitscore** (*float*[145]) – minimum bitscore for accepting a hit

Yields *tuple* – a tuple whose first element is the *uid* (the sequence name) and the second is the a list of tuples whose first element is the GID (NCBI identifier), the second one is the identity and the third is the bitscore of the hit.

mgkit.io.blast.**parse_uniprot_blast**(*file_handle*, *bitscore=40*, *db='UNIPROT-SP'*, *dbq=10*, *name_func=None*, *feat_type='CDS'*, *seq_lengths=None*)

New in version 0.1.12.

Changed in version 0.1.13: added *name_func* argument

Changed in version 0.2.1: added *feat_type*

Changed in version 0.2.3: added *seq_lengths* and added subject *start* and *end* and *e-value*

Parses BLAST results in tabular format using *parse_blast_tab()*, applying a basic bitscore filter. Returns the annotations associated with each BLAST hit.

Parameters

- **file_handle** (*str*[146], *file*) – file name or open file handle
- **bitscore** (*int*[147], *float*[148]) – the minimum bitscore for an annotation to be accepted
- **db** (*str*[149]) – database used
- **dbq** (*int*[150]) – an index indicating the quality of the sequence database used; this value is used in the filtering of annotations

[144] https://docs.python.org/3/library/stdtypes.html#str
[145] https://docs.python.org/3/library/functions.html#float
[146] https://docs.python.org/3/library/stdtypes.html#str
[147] https://docs.python.org/3/library/functions.html#int
[148] https://docs.python.org/3/library/functions.html#float
[149] https://docs.python.org/3/library/stdtypes.html#str
[150] https://docs.python.org/3/library/functions.html#int

- **name_func** (*func*) – function to convert the name of the database sequences. Defaults to *lambda x: x.split('|')[1]*, which can be be used with fasta files provided by Uniprot
- **feat_type** (*str*[151]) – feature type in the GFF
- **seq_lengths** (*dict*[152]) – dictionary with the sequences lengths, used to deduct the frame of the '-' strand

Yields *Annotation* – instances of *mgkit.io.gff.Annotation* instance of each BLAST hit.

mgkit.io.fasta module

Simple fasta parser and a few utility functions

mgkit.io.fasta.**load_fasta**(*f_handle*)

Changed in version 0.1.13: now returns uppercase sequences

Loads a fasta file and returns a generator of tuples in which the first element is the name of the sequence and the second the sequence

Parameters **f_handle** (*str*[153], *file*) – fasta file to open; a file name or a file handle is expected

Yields *tuple* – first element is the sequence name/header, the second element is the sequence

mgkit.io.fasta.**load_fasta_prodigal**(*file_handle*)

New in version 0.3.1.

Reads a Prodigal aminoacid fasta file and yields a dictionary with basic information about the sequences.

Parameters **file_handle** (*str*[154], *file*) – passed to *load_fasta()*

Yields *dict* – dictionary with the information contained in the header, the last of the attributes put into key *attr*, while the rest are transformed to other keys: seq_id, seq, start, end (genomic), strand, ordinal of

mgkit.io.fasta.**load_fasta_rename**(*file_handle*, *name_func=None*)

New in version 0.3.1.

Renames the header of the sequences using *name_func*, which is called on each header. By default, the behaviour is to keep the header to the left of the first space (BLAST behaviour).

mgkit.io.fasta.**split_fasta_file**(*file_handle*, *name_mask*, *num_files*)

New in version 0.1.13.

Splits a fasta file into a series of smaller files.

Parameters

- **file_handle** (*file*, *str*[155]) – fasta file with the input sequences
- **name_mask** (*str*[156]) – file name template for the splitted files, more informations are found in `mgkit.io.split_write()`
- **num_files** (*int*[157]) – number of files in which to distribute the sequences

mgkit.io.fasta.**write_fasta_sequence**(*file_handle*, *name*, *seq*, *wrap=60*, *write_mode='a'*)

Write a fasta sequence to file. If the *file_handle* is a string, the file will be opened using *write_mode*.

[151] https://docs.python.org/3/library/stdtypes.html#str
[152] https://docs.python.org/3/library/stdtypes.html#dict
[153] https://docs.python.org/3/library/stdtypes.html#str
[154] https://docs.python.org/3/library/stdtypes.html#str
[155] https://docs.python.org/3/library/stdtypes.html#str
[156] https://docs.python.org/3/library/stdtypes.html#str
[157] https://docs.python.org/3/library/functions.html#int

Parameters

- **file_handle** – file handle or string.
- **name** (*str*[158]) – header to write for the sequence
- **seq** (*str*[159]) – sequence to write
- **wrap** (*int*[160]) – int for the line wrapping. If None, the sequence will be written in a single line

mgkit.io.fastq module

Fastq utility functions

mgkit.io.fastq.**check_fastq_type**(*qualities*)

Trys to guess the type of quality string used in a Fastq file

Parameters **qualities** (*str*[161]) – string with the quality scores as in the Fastq file

Return str a string with the guessed quality score

Note: Possible values are the following, classified but the values usually used in other softwares:

- ASCII33: sanger, illumina-1.8
- ASCII64: illumina-1.3, illumina-1.5, solexa-old

---

mgkit.io.fastq.**choose_header_type**(*seq_id*)

Return the guessed compiled regular expression :param str seq_id: sequence header to test

Returns compiled regular expression object or None if no match found

mgkit.io.fastq.**convert_seqid_to_new**(*seq_id*)

Convert old seq_id format for Illumina reads to the new found in Casava 1.8+

Parameters **seq_id** (*str*[162]) – seq_id of the sequence (stripped of '@')

Return str the new format seq_id

Note: Example from Wikipedia:

```
old casava seq_id:
@HWUSI-EAS100R:6:73:941:1973#0/1
new casava seq_id:
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCAC
```

---

mgkit.io.fastq.**convert_seqid_to_old**(*seq_id*, *index_as_seq=True*)

Convert old seq_id format for Illumina reads to the new found in Casava until 1.8, which marks the new format.

Parameters

- **seq_id** (*str*[163]) – seq_id of the sequence (stripped of '@')
- **index_as_seq** (*bool*[164]) – if True, the index for the multiplex we'll be the sequence found at the end of the new format seq_id. Otherwise, 0 we'll be used

[158] https://docs.python.org/3/library/stdtypes.html#str
[159] https://docs.python.org/3/library/stdtypes.html#str
[160] https://docs.python.org/3/library/functions.html#int
[161] https://docs.python.org/3/library/stdtypes.html#str
[162] https://docs.python.org/3/library/stdtypes.html#str
[163] https://docs.python.org/3/library/stdtypes.html#str
[164] https://docs.python.org/3/library/functions.html#bool

Return str the new format seq_id

mgkit.io.fastq.**load_fastq**(*file_handle*, *num_qual=False*)

New in version 0.3.1.

Loads a fastq file and returns a generator of tuples in which the first element is the name of the sequence, the second the sequence and the third the quality scores (converted in a numpy array if *num_qual* is True).

Note: this is a simple parser that assumes each sequence is on 4 lines, 1st and 3rd for the headers, 2nd for the sequence and 4th the quality scores

---

Parameters **file_handle** (*str*[165], *file*) – fastq file to open, can be a file name or a file handle

Yields *tuple* – first element is the sequence name/header, the second element is the sequence, the third is the quality score. The quality scores are kept as a string if *num_qual* is False (default) and converted to a numpy array with correct values (0-41) if *num_qual* is True

Raises

- ValueError[166] – if the headers in both sequence and quality scores are not
- valid. This implies that the sequence/qualities have carriage returns
- or the file is truncated.
- TypeError[167] – if the qualities are in a format different than sanger
- (min 0, max 40) or illumina-1.8 (0, 41)

mgkit.io.fastq.**write_fastq_sequence**(*file_handle*, *name*, *seq*, *qual*, *write_mode='a'*)

Write a fastq sequence to file. If the *file_handle* is a string, the file will be opened using *write_mode*.

Parameters

- **file_handle** – file handle or string.
- **name** (*str*[168]) – header to write for the sequence
- **seq** (*str*[169]) – sequence to write
- **qual** (*str*[170]) – quality string

mgkit.io.gff module

This modules define classes and function related to manipulation of GFF/GTF files.

class mgkit.io.gff.**Annotation**(*seq_id='None'*, *start=1*, *end=1*, *strand='+'*, *source='None'*, *feat_type='None'*, *score=0.0*, *phase=0*, *uid=None*, ***kwd*)

Bases: *mgkit.io.gff.GenomicRange*

New in version 0.1.12.

Changed in version 0.2.1: using __slots__ for better memory usage

Alternative implementation for an Annotation. When initialised, If *uid* is None, a unique id is added using *uuid.uuid4*.

[165] https://docs.python.org/3/library/stdtypes.html#str
[166] https://docs.python.org/3/library/exceptions.html#ValueError
[167] https://docs.python.org/3/library/exceptions.html#TypeError
[168] https://docs.python.org/3/library/stdtypes.html#str
[169] https://docs.python.org/3/library/stdtypes.html#str
[170] https://docs.python.org/3/library/stdtypes.html#str

add_exp_syn_count(*seq*, *syn_matrix=None*)
New in version 0.1.13.

Adds expected synonymous/non-synonymous values for an annotation.

Parameters **seq** (*str*[171]) – sequence corresponding to the annotation seq_id syn_matrix (None, dict): matrix that determines the return values. Defaults to the one defined in the called function `mgkit.utils.sequnce.get_seq_expected_syn_count()`.

add_gc_content(*seq*)
Adds GC content information for an annotation. The formula is:

$$\frac{(G+C)}{(G+C+A+T)} \tag{7.1}$$

Modifies the instances of the annotation. gc_ratio will be added to its attributes.

Parameters **seq** (*str*[172]) – nucleotide sequence referred in the GFF

add_gc_ratio(*seq*)
Adds GC content information for an annotation. The formula is:

$$\frac{(A+T)}{(G+C)} \tag{7.2}$$

Modifies the instances of the annotation. gc_ratio will be added to its attributes.

Parameters **seq** (*str*[173]) – nucleotide sequence referred in the GFF

attr

bitscore
bitscore of the annotation

counts
New in version 0.2.2.

Returns the sample counts for the annotation

coverage
New in version 0.1.13.

Return the total coverage for the annotation

Return float coverage

Raises ***AttributeNotFound*** – if no coverage attribute is found

db
db used for the gene_id prediction

dbq
db quality of the annotation

exp_nonsyn
New in version 0.1.13.

Returns the expected number of non-synonymous changes

exp_syn
New in version 0.1.13.

Returns the expected number of synonymous changes

feat_type

[171] https://docs.python.org/3/library/stdtypes.html#str
[172] https://docs.python.org/3/library/stdtypes.html#str
[173] https://docs.python.org/3/library/stdtypes.html#str

fpkms
New in version 0.2.2.

Returns the sample fpkms for the annotation

gene_id
gene_id of the annotation, or *ko* if available

get_aa_seq(*seq*, *start=0*, *tbl=None*, *snp=None*)
New in version 0.1.16.

Returns a translated aminoacid sequence of the annotation. The snp parameter is passed to *Annotation.get_nuc_seq()*

Parameters

- **seq** (*seq*) – chromosome/contig sequence
- **start** (*int*[174]) – position (0-based) from where the correct occurs (frame). If None, the phase attribute is used
- **tbl** (*dict*[175]) – dictionary with the translation for each codon, passed to *mgkit.utils.sequence.translate_sequence()*
- **snp** (*tuple*[176]) – first element is the position of the SNP and the second element is the change

Returns aminoacid sequence

Return type str[177]

get_attr(*attr*, *conv=<type 'str'>*)
New in version 0.1.13.

Generic method to get an attribute and convert it to a specific datatype

get_ec(*level=4*)
New in version 0.1.13.

Changed in version 0.2.0: returns a *set* instead of a list

Returns the EC values associated with the annotation, cutting them at the desired level.

Parameters **level** (*int*[178]) – level of classification desired (between 1 and 4)

Returns list of all EC numbers associated, at the desired level, if none are found an empty set is returned

Return type set[179]

get_mapping(*db*)
New in version 0.1.13.

Returns the mappings, to a particular db, associated with the annotation.

Parameters **db** (*str*[180]) – database to which the mappings come from

Returns list of all mappings associated, to the specified db, if none are found an empty list is returned

Return type list[181]

[174] https://docs.python.org/3/library/functions.html#int
[175] https://docs.python.org/3/library/stdtypes.html#dict
[176] https://docs.python.org/3/library/stdtypes.html#tuple
[177] https://docs.python.org/3/library/stdtypes.html#str
[178] https://docs.python.org/3/library/functions.html#int
[179] https://docs.python.org/3/library/stdtypes.html#set
[180] https://docs.python.org/3/library/stdtypes.html#str
[181] https://docs.python.org/3/library/stdtypes.html#list

get_mappings()
New in version 0.2.1.

Return a dictionary where the keys are the mapping DBs (lowercase) and and the values are the mapping IDs for that DB

get_nuc_seq(*seq*, *reverse=False*, *snp=None*)
New in version 0.1.13.

Changed in version 0.1.16: added *snp* parameter

Returns the nucleotidic sequence that the annotation covers. if the annotation's strand is -, and *reverse* is True, the reverse complement is returned.

Parameters

- **seq** (*seq*) – chromosome/contig sequence
- **reverse** (*bool*[182]) – if True and the strand is '-', a reverse complement is returned
- **snp** (*tuple*[183]) – first element is the position of the SNP relative to the Annotation and the second element is the change

Returns nucleotide sequence with requested transformations

Return type str[184]

get_number_of_samples(*min_cov=4*)
New in version 0.1.13.

Returns the number of sample that have at least a minimum coverage of *min_cov*.

Parameters **min_cov** (*int*[185]) – minimum coverage

Return int number of samples passing the filter

Raises ***AttributeNotFound*** – if no sample coverage attribute is found

is_syn(*seq*, *pos*, *change*, *tbl=None*, *abs_pos=True*, *start=0*)
New in version 0.1.16.

Return if a SNP is synonymous or non-synonymous.

Parameters

- **seq** (*seq*) – reference sequence of the annotation
- **pos** (*int*[186]) – position of the SNP on the reference (1-based index)
- **change** (*str*[187]) – nucleotidic change
- **tbl** (*dict*[188]) – dictionary with the translation table. Defaults to the universal genetic code
- **abs_pos** (*bool*[189]) – if True the *pos* is referred to the reference and not a position relative to the annotation
- **start** (*int*[190] *or None*) – phase to be used to get the start position of the codon. if None, the Annotation phase will be used

Returns True if the SNP is synonymous, false if it's non-synonymous

[182] https://docs.python.org/3/library/functions.html#bool
[183] https://docs.python.org/3/library/stdtypes.html#tuple
[184] https://docs.python.org/3/library/stdtypes.html#str
[185] https://docs.python.org/3/library/functions.html#int
[186] https://docs.python.org/3/library/functions.html#int
[187] https://docs.python.org/3/library/stdtypes.html#str
[188] https://docs.python.org/3/library/stdtypes.html#dict
[189] https://docs.python.org/3/library/functions.html#bool
[190] https://docs.python.org/3/library/functions.html#int

Return type bool[191]

length
Changed in version 0.2.0.

Length of the annotation, uses *len(self)*

phase

region
New in version 0.1.13.

Return the *region* covered by the annotation, to use in samtools

sample_coverage
New in version 0.1.13.

Returns a dictionary with the coverage for each sample, the returned dictionary has the sample id (stripped of the *_cov*) suffix and as values the coverage (converted via `int()`).

Return dict dictionary with the samples' coverage

score

set_attr(*attr*, *value*)
New in version 0.1.13.

Generic method to set an attribute

set_mapping(*db*, *values*)
New in version 0.1.13.

Set mappings to a particular db, associated with the annotation.

Parameters

- **db** (*str*[192]) – database to which the mappings come from
- **mappings** (*iterable*) – iterable of mappings

source

taxon_db
db used for the taxon_id prediction

taxon_id
Changed in version 0.3.1: if taxon_id is set to "None" as a string, it's converted to *None*

taxon_id of the annotation

to_dict(*exclude_attr=None*)
New in version 0.3.1.

Return a dictionary representation of the Annotation.

Parameters **exclude_attr** (*str*[193], *list*[194]) – attributes to exclude from the dictionary, can be either a single attribute (string) or a list of strings

Returns dictionary with the annotation

Return type dict[195]

to_file(*file_handle*)
Writes the GFF annotation to *file_handle*

to_gff(*sep='='*)
Format the Annotation as a GFF string.

[191] https://docs.python.org/3/library/functions.html#bool
[192] https://docs.python.org/3/library/stdtypes.html#str
[193] https://docs.python.org/3/library/stdtypes.html#str
[194] https://docs.python.org/3/library/stdtypes.html#list
[195] https://docs.python.org/3/library/stdtypes.html#dict

Parameters **sep** (*str*[196]) – separator key -> value

Returns annotation formatted as GFF

Return type str[197]

to_gtf(*gene_id_attr='uid'*, *sep=' '*)
New in version 0.1.15.

Changed in version 0.1.16: added *gene_id_attr* parameter

Changed in version 0.2.2: added *sep* argument, default to a space, now

Simple conversion to a valid GTF. gene_id and transcript_id are set to *uid* or the attribute specified using the *gene_id_attr* parameter. It's written to be used with *SNPDat*.

to_json()
New in version 0.2.1.

Changed in version 0.3.1: now *Annotation.to_dict()* is used

Returns a json representation of the Annotation

to_mongodb(*lineage_func=None*, *indent=None*)
New in version 0.2.1.

Changed in version 0.2.2: added handling of *counts_* and *fpkms_*

Changed in version 0.2.6: added *indent* parameter

Returns a MongoDB document that represent the Annotation.

Parameters

- **lineage** (*func*) – function used to populate the lineage key, returns a list of taxon_id
- **indent** (*int*[198]) – the amount of indent to put in the record, None (the default) is for the most compact - one line for the record

Returns the MongoDB document, with Annotation.uid as _id

Return type str[199]

uid
New in version 0.1.13.

uid of the annotation

exception mgkit.io.gff.**AttributeNotFound**
Bases: exceptions.Exception

Raised if an attribute is not found in a GFF file

exception mgkit.io.gff.**DuplicateKeyError**
Bases: exceptions.Exception

New in version 0.1.12.

Raised if a GFF annotation contains duplicate keys

class mgkit.io.gff.**GenomicRange**(*seq_id='None'*, *start=1*, *end=1*, *strand='+'*)
Bases: object[200]

Defines a genomic range

Changed in version 0.2.1: using __slots__ for better memory usage

[196] https://docs.python.org/3/library/stdtypes.html#str
[197] https://docs.python.org/3/library/stdtypes.html#str
[198] https://docs.python.org/3/library/functions.html#int
[199] https://docs.python.org/3/library/stdtypes.html#str
[200] https://docs.python.org/3/library/functions.html#object

__contains__(*pos*)
Changed in version 0.2.3: a range or a subclass are accepted

New in version 0.1.16.

Tests if the position is inside the range of the GenomicRange

Pos is 1-based as *GenomicRange.start* and *GenomicRange.end*

end

expand_from_list(*others*)
Expand the *GenomicRange* range instance with a list of *GenomicRange*

Parameters **others** (*iterable*) – iterable of *GenomicRange*

get_range()
New in version 0.1.13.

Returns the start and end position as a tuple

get_relative_pos(*pos*)
New in version 0.1.16.

Given an absolute position (referred to the reference), convert the position to a coordinate relative to the GenomicRange

Returns the position relative to the GenomicRange

Return type int[201]

Raises ValueError[202] – if the position is not in the range

intersect(*other*)
Return an instance of *GenomicRange* that represent the intersection of the current instance and another.

seq_id

start

strand

union(*other*)
Return the union of two *GenomicRange*

mgkit.io.gff.**annotate_sequence**(*name*, *seq*, *window=None*)

mgkit.io.gff.**annotation_coverage**(*annotations*, *seqs*, *strand=True*)
New in version 0.1.12.

Given a list of annotations and a dictionary where the keys are the sequence names referred in the annotations and the values are the sequences themselves, returns a number which indicated how much the sequence length is "covered" in annotations. If *strand* is True the coverage is strand specific.

Parameters

- **annotations** (*iterable*) – iterable of *Annotation* instances
- **seqs** (*dict*[203]) – dictionary in which the keys are the sequence names and the values are the sequences
- **strand** (*bool*[204]) – if True, the values are strand specific (the annotations) are grouped by (seq_id, strand) instead of seq_id

Yields *tuple* – the first element is the key, (seq_id, strand) if *strand* is True or seq_id if *strand* is False, and the coverage is the second value.

[201] https://docs.python.org/3/library/functions.html#int
[202] https://docs.python.org/3/library/exceptions.html#ValueError
[203] https://docs.python.org/3/library/stdtypes.html#dict
[204] https://docs.python.org/3/library/functions.html#bool

mgkit.io.gff.**annotation_coverage_sorted**(*annotations*, *seqs*, *strand=True*)

New in version 0.3.1.

Given a list of annotations and a dictionary where the keys are the sequence names referred in the annotations and the values are the sequences themselves, returns a number which indicated how much the sequence length is "covered" in annotations. If *strand* is True the coverage is strand specific.

Note: It differs from *annotation_coverage()* because it assumes the annotations are correctly sorted and in the values yielded

---

Parameters

- **annotations** (*iterable*) – iterable of *Annotation* instances
- **seqs** (*dict*[205]) – dictionary in which the keys are the sequence names and the values are the sequences
- **strand** (*bool*[206]) – if True, the values are strand specific (the annotations) are grouped by (seq_id, strand) instead of seq_id

Yields *tuple* – the first element is the seq_id, the second the strand (if strand is True, else it's set to *None*), and the third element is the coverage.

mgkit.io.gff.**annotation_elongation**(*ann1*, *annotations*)

New in version 0.1.12.

Given an *Annotation* instance and a list of the instances of the same class, returns the longest overlapping range that can be found and the annotations that are included in it.

Warning: annotations are not checked for seq_id and strand

Parameters

- **ann1** (Annotation) – annotation to elongate
- **annotations** (*iterable*) – iterable of *Annotation* instances

Returns the first element is the longest range found, while the the second element is a set with the annotations used

Return type tuple[207]

mgkit.io.gff.**convert_gff_to_gtf**(*file_in*, *file_out*, *gene_id_attr='uid'*)

New in version 0.1.16.

Function that uses *Annotation.to_gtf()* to convert a GFF into GTF.

Parameters

- **file_in** (*str*[208], *file*) – either file name or file handle of a GFF file
- **file_out** (*str*[209]) – file name to which write the converted annotations

mgkit.io.gff.**diff_gff**(*files*, *key_func=None*)

New in version 0.1.12.

Returns a simple diff made between a list of gff files. The annotations are grouped using *key_func*, so it depends on it to find similar annotations.

[205] https://docs.python.org/3/library/stdtypes.html#dict
[206] https://docs.python.org/3/library/functions.html#bool
[207] https://docs.python.org/3/library/stdtypes.html#tuple
[208] https://docs.python.org/3/library/stdtypes.html#str
[209] https://docs.python.org/3/library/stdtypes.html#str

Parameters

- **files** (*iterable*) – an iterable of file handles, pointing to GFF files
- **key_func** (*func*) – function used to group annotations, defaults to this key: *(x.seq_id, x.strand, x.start, x.end, x.gene_id, x.bitscore)*

Returns the returned dictionary keys are determined by key_func and as values lists. The lists elements are tuple whose first element is the index of the file, relative to *files* and the second element is the line number in which the annotation is. Can be used with the `linecache`[210] module.

Return type dict[211]

mgkit.io.gff.**elongate_annotations**(*annotations*)

New in version 0.1.12.

Given an iterable of *Annotation* instances, tries to find the all possible longest ranges and returns them.

> **Warning:** annotations are not checked for seq_id and strand

Parameters **annotations** (*iterable*) – iterable of *Annotation* instances

Returns set with the all ranges found

Return type set[212]

mgkit.io.gff.**extract_nuc_seqs**(*annotations*, *seqs*, *name_func=<function <lambda>>*, *reverse=False*)

New in version 0.1.13.

Extract the nucleotidic sequences from a list of annotations. Internally uses the method *Annotation.get_nuc_seq()*.

Parameters

- **annotations** (*iterable*) – iterable of *Annotation* instances
- **seqs** (*dict*[213]) – dictionary with the sequences referenced in the annotations
- **name_func** (*func*) – function used to extract the sequence name to be used, defaults to the uid of the annotation
- **reverse** (*bool*[214]) – if True the annotations on the - strand are reverse complemented

Yields *tuple* – tuple whose first element is the sequence name and the second is the sequence to which the annotation refers.

mgkit.io.gff.**from_aa_blast_frag**(*hit*, *parent_ann*, *aa_seqs*)

mgkit.io.gff.**from_gff**(*line*, *strict=True*)

New in version 0.1.12.

Changed in version 0.2.6: added *strict* parameter

Parse GFF line and returns an *Annotation* instance

Parameters

- **line** (*str*[215]) – GFF line
- **strict** (*bool*[216]) – if True duplicate keys raise an exception

[210] https://docs.python.org/3/library/linecache.html#module-linecache
[211] https://docs.python.org/3/library/stdtypes.html#dict
[212] https://docs.python.org/3/library/stdtypes.html#set
[213] https://docs.python.org/3/library/stdtypes.html#dict
[214] https://docs.python.org/3/library/functions.html#bool
[215] https://docs.python.org/3/library/stdtypes.html#str
[216] https://docs.python.org/3/library/functions.html#bool

Returns instance of *Annotation* for the line

Return type *Annotation*

Raises *DuplicateKeyError* – if the attribute column has duplicate keys

mgkit.io.gff.**from_glimmer3**(*header*, *line*, *feat_type='CDS'*)

New in version 0.1.12.

Parses the line of a GLIMMER3 ouput and returns an instance of a GFF annotation.

Parameters

- **header** (*str*[217]) – the seq_id to which the ORF belongs
- **line** (*str*[218]) – the prediction line for the orf
- **feat_type** (*str*[219]) – the feature type to use

Returns instance of annotation

Return type *Annotation*

Example

Assuming a GLIMMER3 output like this:

```
>sequence0001
orf00001        66       611  +3      6.08
```

The code used is:

```
>>> header = 'sequence0001'
>>> line = 'orf00001        66       611  +3      6.08'
>>> from_glimmer3(header, line)
```

mgkit.io.gff.**from_hmmer**(*line*, *aa_seqs*, *feat_type='gene'*, *source='HMMER'*, *db='CUSTOM'*, *custom_profiles=True*, *noframe=False*)

New in version 0.1.15: first implementation to move old scripts to new GFF specs

Changed in version 0.2.1: removed compatibility with old scripts

Changed in version 0.2.2: taxon_id and taxon_name are not saved for non-custom profiles

Changed in version 0.3.1: added support for non mgkit-translated sequences (*noframe*)

Parse HMMER results (one line), it won't parse commented lines (starting with #)

Parameters

- **line** (*str*[220]) – HMMER domain table line
- **aa_seqs** (*dict*[221]) – dictionary with amino-acid sequences (name->seq), used to get the correct nucleotide positions
- **feat_type** (*str*[222]) – string to be used in the 'feature type' column
- **source** (*str*[223]) – string to be used in the 'source' column
- **custom_profiles** (*bool*[224]) – if True, the profile name contains gene, taxonomy and reviewed information in the form KOID_TAXONID_TAXON-NAME(-nr)

[217] https://docs.python.org/3/library/stdtypes.html#str
[218] https://docs.python.org/3/library/stdtypes.html#str
[219] https://docs.python.org/3/library/stdtypes.html#str
[220] https://docs.python.org/3/library/stdtypes.html#str
[221] https://docs.python.org/3/library/stdtypes.html#dict
[222] https://docs.python.org/3/library/stdtypes.html#str
[223] https://docs.python.org/3/library/stdtypes.html#str
[224] https://docs.python.org/3/library/functions.html#bool

- **noframe** (*bool*[225]) – if True, the sequence is assumed to be in frame f0

Returns A *Annotation* instance

Note: if *custom_profiles* is False, gene_id, taxon_id and taxon_name will be equal to the profile name

---

mgkit.io.gff.**from_json**(*line*)
New in version 0.2.1.

Returns an Annotation from a json representation

mgkit.io.gff.**from_mongodb**(*record*, *lineage=True*)
New in version 0.2.1.

Changed in version 0.2.2: added handling of *counts_* and *fpkms_*

Changed in version 0.2.6: better handling of missing attributes and added *lineage* parameter

Returns a *Annotation* instance from a MongoDB record (created) using *Annotation.to_mongodb()*. The actual record returned by pymongo is a dictionary that is copied, manipulated and passed to the `Annotation.__init__()`.

Parameters

- **record** (*dict*[226]) – a dictionary with the full record from a MongoDB query
- **lineage** (*bool*[227]) – indicates if the lineage information in the record should be kept in the annotation

Returns instance of *Annotation* object

Return type *Annotation*

mgkit.io.gff.**from_nuc_blast**(*hit*, *db*, *feat_type='CDS'*, *seq_len=None*, *to_nuc=False*, ***kwd*)
New in version 0.1.12.

Changed in version 0.1.16: added *to_nuc* parameter

Changed in version 0.2.3: removed *to_nuc*, the hit can include the subject end/start and evalue

Returns an instance of *Annotation*

Parameters

- **hit** (*tuple*[228]) – a BLAST hit, from *mgkit.io.blast.parse_blast_tab()*
- **db** (*str*[229]) – db used with BLAST

Keyword Arguments

- **feat_type** (*str*[230]) – feature type in the GFF
- **seq_len** (*int*[231]) – sequence length, if supplied, the phase for strand '-' can be assigned, otherwise is assigned a 0
- ****kwd** – any additional column

Returns instance of *Annotation*

Return type *Annotation*

mgkit.io.gff.**from_nuc_blast_frag**(*hit*, *parent_ann*, *db='NCBI-NT'*)

[225] https://docs.python.org/3/library/functions.html#bool
[226] https://docs.python.org/3/library/stdtypes.html#dict
[227] https://docs.python.org/3/library/functions.html#bool
[228] https://docs.python.org/3/library/stdtypes.html#tuple
[229] https://docs.python.org/3/library/stdtypes.html#str
[230] https://docs.python.org/3/library/stdtypes.html#str
[231] https://docs.python.org/3/library/functions.html#int

mgkit.io.gff.**from_prodigal_frag**(*main_gff*, *blast_gff*, *attr='ID'*, *split_func=None*)
New in version 0.2.6: *experimental*

Reads the GFF given in output by PRODIGAL and the resulting GFF from using BLAST (or other software) on the aa or nucleotide file output by PRODIGAL.

It then integrates the two outputs, so to the PRODIGAL GFF is added the information from the the output of the gene prediction software used.

Parameters

- **main_gff** (*file*) – GFF file from PRODIGAL
- **blast_gff** (*file*) – GFF with the returned annotations
- **attr** (*str*[232]) – attribute in the PRODIGAL GFF that is used to identify an annotation
- **split_func** (*func*) – function to rename the headers from the predicted sequences back to their parent sequence

Yields *annotation* – annotation for each *blast_gff* back translated

mgkit.io.gff.**from_sequence**(*name*, *seq*, *feat_type='SEQUENCE'*, ***kwd*)
New in version 0.1.12.

Returns an instance of *Annotation* for the full length of a sequence

Parameters

- **name** (*str*[233]) – name of the sequence
- **seq** (*str*[234]) – sequence, to get the length of the annotation

Keyword Arguments

- **feat_type** (*str*[235]) – feature type in the GFF
- ****kwd** – any additional column

Returns instance of *Annotation*

Return type *Annotation*

mgkit.io.gff.**get_annotation_map**(*annotations*, *key_func*, *value_func*)
New in version 0.1.15.

Applies two functions to an iterable of annotations with an iterator returned with the applied functions. Useful to build a dictionary

Parameters

- **annotations** (*iterable*) – iterable of annotations
- **key_func** (*func*) – function that accept an annotation as argument and returns one value, the first of the returned tuple
- **value_func** (*func*) – function that accept an annotation as argument and returns one value, the second of the returned tuple

Yields *tuple* – a tuple where the first value is the result of *key_func* on the passed annotation and the second is the value returned by *value_func* on the same annotation

mgkit.io.gff.**group_annotations**(*annotations*, *key_func=<function <lambda>>*)
New in version 0.1.12.

Group *Annotation* instances in a dictionary by using a key function that returns the key to be used in the dictionary.

[232] https://docs.python.org/3/library/stdtypes.html#str
[233] https://docs.python.org/3/library/stdtypes.html#str
[234] https://docs.python.org/3/library/stdtypes.html#str
[235] https://docs.python.org/3/library/stdtypes.html#str

Parameters

- **annotations** (*iterable*) – iterable with *Annotation* instances
- **key_func** (*func*) – function used to extract the key used in the dictionary, defaults to a function that returns (ann.seq_id, ann.strand)

Returns dictionary whose keys are returned by *key_func* and the values are lists of annotations

Return type dict[236]

Example

```
>>> ann = [Annotation(seq_id='seq1', strand='+', start=10, end=15),
... Annotation(seq_id='seq1', strand='+', start=1, end=5),
... Annotation(seq_id='seq1', strand='-', start=30, end=100)]
>>> group_annotations(ann)
{('seq1', '+'): [seq1(+):10-15, seq1(+):1-5], ('seq1', '-'): [seq1(-):30-100]}
```

mgkit.io.gff.**group_annotations_by_ancestor**(*annotations*, *ancestors*, *taxonomy*)

New in version 0.1.13.

Group annotations by the ancestors provided.

Parameters

- **annotations** (*iterable*) – annotations to group
- **ancestors** (*iterable*) – list of ancestors accepted
- **taxonomy** – taxonomy class

Returns grouped annotations

Return type dict[237]

mgkit.io.gff.**group_annotations_sorted**(*annotations*, *key_func=<function <lambda>>*)

New in version 0.1.13.

Group *Annotation* instances by using a key function that returns a key. Assumes that the annotations are already sorted to return an iterator and save memory. One way to sort them is using: *sort -s -k 1,1 -k 7,7* on the file.

Parameters

- **annotations** (*iterable*) – iterable with *Annotation* instances
- **key_func** (*func*) – function used to extract the key used in the dictionary, defaults to a function that returns (ann.seq_id, ann.strand)

Yields *list* – a list of the grouped annotations by *key_func* values

mgkit.io.gff.**load_gff_base_info**(*files*, *taxonomy=None*, *exclude_ids=None*, *include_taxa=None*)

This function is useful if the number of annotations in a GFF is high or there are memory constraints on the system. It returns a dictionary that can be used with functions like *mgkit.counts.func.load_sample_counts()*.

Parameters

- **files** (*iterable, str*[238]) – file name or list of paths of GFF files
- **taxonomy** – taxonomy pickle file, needed if include_taxa is not None

[236] https://docs.python.org/3/library/stdtypes.html#dict
[237] https://docs.python.org/3/library/stdtypes.html#dict
[238] https://docs.python.org/3/library/stdtypes.html#str

- **exclude_ids** (*set*[239], *list*[240]) – a list of gene_id to exclude from the dictionary
- **include_taxa** (*int*[241], *iterable*) – a taxon_id or list thereof to be passed to `mgkit.taxon.taxonomy.is_ancestor()`, so only the taxa that have the those taxon_id(s) as ancestor(s) are kept

Returns dictionary where the key is *Annotation.uid* and the value is a tuple (*Annotation.gene_id*, *Annotation.taxon_id*)

Return type dict[242]

`mgkit.io.gff.`**load_gff_mappings**(*files*, *map_db*, *taxonomy=None*, *exclude_ids=None*, *include_taxa=None*)

This function is useful if the number of annotations in a GFF is high or there are memory constraints on the system. It returns a dictionary that can be used with functions like *mgkit.counts.func.load_sample_counts()*.

Parameters

- **files** (*iterable*, *str*[243]) – file name or list of paths of GFF files
- **map_db** (*str*[244]) – any kind mapping in the GFF, as passed to *Annotation.get_mapping()*
- **taxonomy** – taxonomy pickle file, needed if include_taxa is not None
- **exclude_ids** (*set*[245], *list*[246]) – a list of gene_id to exclude from the dictionary
- **include_taxa** (*int*[247], *iterable*) – a taxon_id or list thereof to be passed to `mgkit.taxon.taxonomy.is_ancestor()`, so only the taxa that have the those taxon_id(s) as ancestor(s) are kept

Returns dictionary where the key is *Annotation.gene_id* and the value is a list of mappings, as returned by *Annotation.get_mapping()*

Return type dict[248]

`mgkit.io.gff.`**parse_gff**(*file_handle*, *gff_type=<function from_gff>*, *strict=True*)

Changed in version 0.2.6: added *strict* parameter

Changed in version 0.2.3: correctly handling of GFF with comments of appended sequences

Changed in version 0.1.12: added *gff_type* parameter

Parse a GFF file and returns generator of `GFFKegg` instances

Accepts a file handle or a string with the file name

Parameters

- **file_handle** (*str*[249], *file*) – file name or file handle to read from
- **gff_type** (*class*) – class/function used to parse a GFF annotation
- **strict** (*bool*[250]) – if True duplicate keys raise an exception

Yields *Annotation* – an iterator of *Annotation* instances

[239] https://docs.python.org/3/library/stdtypes.html#set
[240] https://docs.python.org/3/library/stdtypes.html#list
[241] https://docs.python.org/3/library/functions.html#int
[242] https://docs.python.org/3/library/stdtypes.html#dict
[243] https://docs.python.org/3/library/stdtypes.html#str
[244] https://docs.python.org/3/library/stdtypes.html#str
[245] https://docs.python.org/3/library/stdtypes.html#set
[246] https://docs.python.org/3/library/stdtypes.html#list
[247] https://docs.python.org/3/library/functions.html#int
[248] https://docs.python.org/3/library/stdtypes.html#dict
[249] https://docs.python.org/3/library/stdtypes.html#str
[250] https://docs.python.org/3/library/functions.html#bool

mgkit.io.gff.**parse_gff_files**(*files*, *strict=True*)

New in version 0.1.15.

Changed in version 0.2.6: added *strict* parameter

Function that returns an iterator of annotations from multiple GFF files.

Parameters

- **files** (*iterable, str*[251]) – iterable of file names of GFF files, or a single file name
- **strict** (*bool*[252]) – if True duplicate keys raise an exception

Yields *Annotation* – iterator of annotations

mgkit.io.gff.**split_gff_file**(*file_handle*, *name_mask*, *num_files=2*)

New in version 0.1.14.

Changed in version 0.2.6: now accept a file object as sole input

Splits a GFF, or a list of them, into a number of files. It is assured that annotations for the same sequence are kept in the same file, which is useful for cases like filtering, even when the annotations are from different GFF files.

Internally, a structure is kept to check if a sequence ID is already been stored to a file, in which case the annotation is written to that file, otherwise a random file handles (among the open ones) is chosen.

Parameters

- **file_handle** (*str*[253], *list*[254]) – a single or list of file handles (or file names), from which the GFF annotations are read
- **name_mask** (*str*[255]) – a string used as template for the output file names on which the function applies `string.format()`
- **num_files** (*int*[256]) – the number of files to split the records

Example

```
>>> import glob
>>> files = glob.glob('*.gff')
>>> name_mask = 'split-file-{0}.gff'
>>> split_gff_file(files, name_mask, 5)
```

mgkit.io.gff.**write_gff**(*annotations*, *file_handle*, *verbose=True*)

Changed in version 0.1.12: added *verbose* argument

Write a GFF to file

Parameters

- **annotations** (*iterable*) – iterable that returns `GFFKegg` or *Annotation* instances
- **file_handle** (*str*[257], *file*) – file name or file handle to write to
- **verbose** (*bool*[258]) – if True, a message is logged

[251] https://docs.python.org/3/library/stdtypes.html#str
[252] https://docs.python.org/3/library/functions.html#bool
[253] https://docs.python.org/3/library/stdtypes.html#str
[254] https://docs.python.org/3/library/stdtypes.html#list
[255] https://docs.python.org/3/library/stdtypes.html#str
[256] https://docs.python.org/3/library/functions.html#int
[257] https://docs.python.org/3/library/stdtypes.html#str
[258] https://docs.python.org/3/library/functions.html#bool

mgkit.io.glimmer module

mgkit.io.glimmer.**parse_glimmer3**(*file_handle*)

Parses an ouput file from glimmer3 and yields the header and prediction lines. Used to feed the *mgkit.io.gff.from_glimmer3()* function.

Parameters **file_handle** (*str*[259], *file*) – file name or file handle to read from

Yields *tuple* – first element is the sequence of the predicted gene and the second is the prediction line

mgkit.io.snpdat module

SNPDat reader

class mgkit.io.snpdat.**SNPDatRow**(*line=None*, *rev_comp=None*)

Bases: object[260]

Class containing information ouputted by SNPDat in its result file. One instance contains information about a row in the file.

chr_name
str – the queried SNPs chromosome ID

chr_pos
int – queried SNPs genomic location

in_feat
bool – Whether or not the queried SNP was within a feature

region
str – Region containing the SNP; either exonic, intronic or intergenic

feat_dist
int – Distance to nearest feature

feature
str – Either the closest feature to the SNP or the feature containing the SNP

num_features
int – number of different features that the SNP is annotated to

feat_num
int – number of annotations of the current feature

feat_start
int – Start of feature (bp)

feat_end
int – End of feature (bp)

gene_id
str – gene ID for the current feature

gene_name
str – gene name for the current feature

transcript_id
str – transcript ID for the current feature

transcript_name
str – transcript name for the current feature

[259] https://docs.python.org/3/library/stdtypes.html#str
[260] https://docs.python.org/3/library/functions.html#object

exon
tuple – exon that contains the current feature and the total number of annotated exons for the gene containing the feature

strand
str – strand sense of the feature

ann_frame
str – annotated reading frame (when contained in the GTF)

frame
str – reading frame estimated by SNPdat

num_stops
int – estimated number of stop codons in the estimated reading frame

codon
tuple – codon containing the SNP, position in the codon and reference base and mutation

nuc_change
tuple – amino acid for the reference codon and new amino acid with the mutation in place

nuc_ref
str, None – reference nucleotide

aa_change
str – amino acid for the reference codon and new amino acid with the mutation in place

synonymous
bool – Whether or not the mutation is synonymous

protein_id
str – protein ID for the current feature

messages
str – messages in the SNPDat line

aa_change

ann_frame

chr_name

chr_pos

codon

exon

feat_dist

feat_end

feat_num

feat_start

feature

frame

gene_id

gene_name

in_feat

messages

nuc_change

nuc_ref

num_features

num_stops

protein_id

region

strand

synonymous

transcript_id

transcript_name

mgkit.io.snpdat.**snpdat_reader**(*f_handle*)
Simple SNPDat reader.

f_handle: file handle or string for the SNPDat result file

Returns generator of SNPDatRow instances

mgkit.io.uniprot module

New in version 0.1.13.

Uniprot file formats

mgkit.io.uniprot.**parse_uniprot_mappings**(*file_handle*, *gene_ids=None*, *mappings=None*, *num_lines=10000000*)
Parses a Uniprot mapping file[261], returning a generator with the mappings.

Parameters

- **file_handle** (*str*[262], *file*) – file name or open file handle
- **gene_ids** (*None*, *set*[263]) – if not None, the returned mappings are for the gene IDs specified
- **mappings** (*None*, *set*[264]) – mappings to be returned
- **num_lines** (*None*, *int*[265]) – number of which a message is logged. If None, no message is logged

Yields *tuple* – the first element is the gene ID, the second is the mapping type and third element is the mapped ID

mgkit.io.uniprot.**uniprot_mappings_to_dict**(*file_handle*, *gene_ids*, *mappings*)
Parses a Uniprot mapping file[266], returning a generator of dictionaries with the mappings requested.

Parameters

- **file_handle** (*str*[267], *file*) – file name or open file handle
- **gene_ids** (*None*, *set*[268]) – if not None, the returned mappings are for the gene IDs specified
- **mappings** (*None*, *set*[269]) – mappings to be returned

[261] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping.dat.gz
[262] https://docs.python.org/3/library/stdtypes.html#str
[263] https://docs.python.org/3/library/stdtypes.html#set
[264] https://docs.python.org/3/library/stdtypes.html#set
[265] https://docs.python.org/3/library/functions.html#int
[266] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping.dat.gz
[267] https://docs.python.org/3/library/stdtypes.html#str
[268] https://docs.python.org/3/library/stdtypes.html#set
[269] https://docs.python.org/3/library/stdtypes.html#set

Yields *tuple* – the first element is the gene ID, the second is a dictionary with all the mappings found, the key is the mapping type and the value is a list of all mapped IDs

mgkit.io.utils module

Various utilities to help read and process files

exception mgkit.io.utils.**UnsupportedFormat**
Bases: exceptions.IOError

Raised if the a file can't be opened with the correct module

mgkit.io.utils.**compressed_handle**(*file_handle*)
New in version 0.1.13.

Tries to wrap a file handle in the appropriate compressed file class.

Parameters **file_handle** (*str*[270]) – file handle

Returns the same file handle if no suitable compressed file class is found or the new file_handle which supports the compression

Return type file

Raises *UnsupportedFormat* – if the module to open the file is not available

mgkit.io.utils.**group_tuples_by_key**(*iterator*, *key_func=None*, *skip_elements=0*)
New in version 0.3.1.

Group the elements of an iterator by a key and yields the grouped elements. The elements yielded by the iterator are assumed to be a list or tuple, with the default key (when *key_func* is None) being the first of the of the objects inside that element. This behaviour can be customised by passing to *key_func* a function that accept an element and returns the key to be used.

Note: the iterable assumen that the elements are already sorted by their keys

Parameters

- **iterator** (*iterable*) – iterator to be grouped
- **key_func** (*func*) – function that accepts a element and returns its associated key
- **skip_elements** (*int*[271]) – number of elements to skip at the start

Yields *list* – a list of the grouped elements by key

mgkit.io.utils.**open_file**(*file_name*, *mode='r'*)
New in version 0.1.12.

Opens a file using the extension as a guide to which module to use.

Parameters

- **file_name** (*str*[272]) – file name
- **mode** (*str*[273]) – mode used to open the file

Returns file handle

Return type file

Raises *UnsupportedFormat* – if the module to open the file is not available

[270] https://docs.python.org/3/library/stdtypes.html#str
[271] https://docs.python.org/3/library/functions.html#int
[272] https://docs.python.org/3/library/stdtypes.html#str
[273] https://docs.python.org/3/library/stdtypes.html#str

mgkit.io.utils.**split_write**(*records*, *name_mask*, *write_func*, *num_files=2*)

New in version 0.1.13.

Splits the writing of a number of records in a series of files. The *name_mask* is used as template for the file names. A string like "split-files-{0}" can be specified and the function applies format with the index of the pieces.

Parameters

- **records** (*iterable*) – an iterable that returns a object to be saved
- **name_mask** (*str*[274]) – a string used as template for the output file names on which the function applies `string.format()`
- **write_func** (*func*) – a function that is called to write to the files. It needs to accept a file handles as first argument and the record returned by *records* as the second argument
- **num_files** (*int*[275]) – the number of files to split the records

Module contents

Package used to contain code related to I/O operations

mgkit.mappings package

Submodules

mgkit.mappings.cazy module

Module containing classes and functions to deal with CaZy data

class mgkit.mappings.cazy.**Kegg2CazyMapper**(*fname=None*)

Bases: *mgkit.kegg.KeggMapperBase*

Class holding mappings KOs->CaZy

columns_string = 'database(cazy)'

map_kos_cazy(*kos*, *contact*, *skip=False*)

Map a list of KOs to CaZy ids

query_string = 'database:(type:ko {0}) AND database:(type:cazy)'

mgkit.mappings.cazy.**download_data**(*contact*, *kegg_data='kegg.pickle'*, *cazy_data='cazy.pickle'*)

Function to download CaZy data

mgkit.mappings.eggnog module

Module containing classes and functions to deal with eggNOG data

Todo:

- unify download of data from web

[274] https://docs.python.org/3/library/stdtypes.html#str
[275] https://docs.python.org/3/library/functions.html#int

class mgkit.mappings.eggnog.**Kegg2NogMapper**(*fname=None*, *gen_ko_to_cat=True*)
Bases: *mgkit.kegg.KeggMapperBase*

Usage

to add a ko_id to the mapper use: * map_ko_to_eggnog(ko_id) which calls:

- get_uniprot_from_ko
- get_eggnog_from_uniprot

return None if there was no problem or a dictionary with the exception thrown

to make mappings ko_id -> eggnog-categories: * get_cat_mapping_from_file (file are in the current directory) * gen_ko_to_cat

---

Todo:

- add method to load kos from file
- optimise querying (like Kegg2CazyMapper)

---

columns_string = 'database(eggnog)'

gen_ko_to_cat()

get_cat_ko_dict()

get_cat_mapping(*f_handle*)

get_cat_mapping_from_file(*files=('COG.funccat.txt', 'NOG.funccat.txt', 'KOG.funccat.txt')*)
Duplicate functionality of load_func_cat

Categories are single letter and their descriptioons are in EGGNOG_CAT, while the broader categories EGGNOG_CAT_MAP

get_ko_cat(*ko_id*)

get_ko_cat_dict()

get_ko_map()
Returns a copy of the KO->mapping dictionary

load_data(*fname*)
Loads mapping data to disk

map_ko_to_eggnog(*ko_id*, *contact*)

map_kos_eggnog(*kos*, *contact*)

query_string = 'database:(type:ko {0}) AND database:(type:eggnog)'

save_data(*fname*)
Saves mapping data to disk

class mgkit.mappings.eggnog.**NOGInfo**
Bases: object[276]

New in version 0.1.14.

Mappings from Uniprot to eggNOG

..note:

```
load_description is optional
```

get_gene_funccat(*gene_id*)
Returns the functional category (one letter, EGGNOG_CAT keys) for the requested eggNOG gene ID

[276] https://docs.python.org/3/library/functions.html#object

get_gene_nog(*gene_id*)
Returns the COG/NOG ID of the requested eggNOG gene ID

get_nog_funccat(*nog_id*)
Returns the functional category (one letter, EGGNOG_CAT keys) for the requested eggNOG COG/NOG ID

get_nog_gencat(*nog_id*)
Returns the functional category (EGGNOG_CAT_NAMES keys) for the requested eggNOG COG/NOG ID

get_nogs_funccat(*nog_ids*)
Returns the functional categories for a list of COG/NOG IDs. Uses *NOGInfo.get_nog_funccat()*

load_description(*file_handle*)
Loads data from *NOG.description.txt.gz*

file_handle can either an open file or a path

load_funccat(*file_handle*)
Loads data from *NOG.funccat.txt.gz*

file_handle can either an open file or a path

load_members(*file_handle*)
Loads data from *NOG.members.txt.gz*

file_handle can either an open file or a path

mgkit.mappings.eggnog.**download_data**(*contact*, *kegg_data='kegg.pickle'*, *eggnog_data='eggnog.pickle'*, *base_url='http://eggnog.embl.de/version_3.0/data/downloads/'*)

mgkit.mappings.eggnog.**get_general_eggnog_cat**(*category*)
New in version 0.1.14.

Returns the functional category (EGGNOG_CAT_NAMES keys) for the requested single letter functional category (EGGNOG_CAT keys)

mgkit.mappings.eggnog.**print_ko_to_cat**(*data_file='eggnog.pickle'*, *descr=False*)

mgkit.mappings.enzyme module

New in version 0.1.14.

EC mappings

mgkit.mappings.enzyme.**change_mapping_level**(*ec_map*, *level=3*)
New in version 0.1.14.

Given a dictionary, whose values are dictionaries, in which a key is named *ec* and its value is an iterable of EC numbers, returns an iterator that can be used to build a dictionary with the same top level keys and the values are sets of the transformed EC numbers.

Parameters

- **ec_map** (*dict*[277]) – dictionary generated by *mgkit.net.uniprot.get_gene_info()*
- **level** (*int*[278]) – number from 1 to 4, to specify the level of the mapping, passed to *get_enzyme_level()*

Yields *tuple* – a tuple (gene_id, set(ECs)), which can be passed to *dict* to make a dictionary

[277] https://docs.python.org/3/library/stdtypes.html#dict
[278] https://docs.python.org/3/library/functions.html#int

Example

```
>>> from mgkit.net.uniprot import get_gene_info
>>> from mgkit.mappings.enzyme import change_mapping_level
>>> ec_map = get_gene_info('Q9HFQ1', columns='ec')
{'Q9HFQ1': {'ec': '1.1.3.4'}}
>>> dict(change_mapping_level(ec_map, level=2))
{'Q9HFQ1': {'1.1'}}
```

mgkit.mappings.enzyme.**get_enzyme_full_name**(*ec_id*, *ec_names*, *sep=', '*)

New in version 0.2.1.

From a EC identifiers and a dictionary of names builds a comma separated name (by default) that identifies the function of the enzyme.

Parameters

- **ec_id** (*str*[279]) – EC identifier
- **ec_names** (*dict*[280]) – a dictionary of names that can be produced using *parse_expasy_file()*
- **sep** (*str*[281]) – string used to join the names

Returns the enzyme classification name

Return type str[282]

mgkit.mappings.enzyme.**get_enzyme_level**(*ec*, *level=4*)

New in version 0.1.14.

Returns an enzyme class at a specific level , between 1 and 4 (by default the most specific, 4)

Parameters

- **ec** (*str*[283]) – a string representing an EC number (e.g. 1.2.4.10)
- **level** (*int*[284]) – from 1 to 4, to get a different level specificity of in the enzyme classification

Returns the EC number at the requested specificity

Return type str[285]

Example

```
>>> from mgkit.mappings.enzyme import get_enzyme_level
>>> get_enzyme_level('1.1.3.4', 1)
'1'
>>> get_enzyme_level('1.1.3.4', 2)
'1.1'
>>> get_enzyme_level('1.1.3.4', 3)
'1.1.3'
>>> get_enzyme_level('1.1.3.4', 4)
'1.1.3.4'
```

mgkit.mappings.enzyme.**get_mapping_level**(*ec_map*, *level=3*)

New in version 0.3.0.

[279] https://docs.python.org/3/library/stdtypes.html#str
[280] https://docs.python.org/3/library/stdtypes.html#dict
[281] https://docs.python.org/3/library/stdtypes.html#str
[282] https://docs.python.org/3/library/stdtypes.html#str
[283] https://docs.python.org/3/library/stdtypes.html#str
[284] https://docs.python.org/3/library/functions.html#int
[285] https://docs.python.org/3/library/stdtypes.html#str

Given a dictionary, whose values are iterable of EC numbers, returns an iterator that can be used to build a dictionary with the same top level keys and the values are sets of the transformed EC numbers.

Parameters

- **ec_map** (*dict*[286]) – dictionary genes to EC
- **level** (*int*[287]) – number from 1 to 4, to specify the level of the mapping, passed to *get_enzyme_level()*

Yields *tuple* – a tuple (gene_id, set(ECs)), which can be passed to *dict* to make a dictionary

mgkit.mappings.enzyme.**parse_expasy_file**(*file_name*)
Used to load enzyme descriptions from the file *enzclass.txt* on expasy[288].

The FTP url for enzclass.txt is: ftp://ftp.expasy.org/databases/enzyme/enzclass.txt

mgkit.mappings.go module

Module containing classes and functions to deal with Gene Ontology data

class mgkit.mappings.go.**Kegg2GOMapper**(*fname=None*)
Bases: *mgkit.kegg.KeggMapperBase*

Class holding mappings KOs->GO

columns_string = 'go-id'

load_go_names(*fname*)

load_goslim_mapping(*f_handle*)

map_kos_go(*kos*, *contact*, *skip=False*)
Map a list of KOs to GO ids

query_string = 'database:(type:ko {0})'

write_association_file(*f_handle*)

mgkit.mappings.go.**download_data**(*contact*, *kegg_data='kegg.pickle'*, *go_data='go_data.pickle'*)
Function to download Gene Ontology data

mgkit.mappings.pandas_map module

Module that contains mapping operations on pandas data structures

mgkit.mappings.pandas_map.**calc_coefficient_of_variation**(*dataframe*)
Calculate coefficient of variation for a DataFrame. Uses formula from Wikipedia[289]

The formula used is $\left(1 + \frac{1}{4n}\right) * c_v$ where $c_v = \frac{s}{\bar{x}}$

mgkit.mappings.pandas_map.**concatenate_and_rename_tables**(*dataframes*, *roots*)
Concatenates a list of pandas.DataFrame instances and renames the columns prepending a string to each column in each table from a list of prefixes.

Parameters

- **dataframes** (*iterable*) – iterable of DataFrame instances
- **roots** (*iterable*) – list of prefixes to append to the column names of each DataFrame

[286] https://docs.python.org/3/library/stdtypes.html#dict
[287] https://docs.python.org/3/library/functions.html#int
[288] http://expasy.org
[289] http://en.wikipedia.org/wiki/Coefficient_of_variation

Return DataFrame returns a DataFrame instance

Todo:

- move to pandas_utils?

---

mgkit.mappings.pandas_map.**group_dataframe_by_mapping**(*dataframe*, *mapping*, *root_taxon*, *name_dict=None*)

Return a `pandas.DataFrame` filtered by mapping and root taxon, the values for each column is averaged over all genes mapping to a category.

Parameters

- **dataframe** (*DataFrame*) – DataFrame with multindex gene-root
- **mapping** (*dict*[290]) – dictionary of category->genes
- **root_taxon** (*str*[291]) – root taxon to group genes
- **name_dict** (*dict*[292]) – dictionary of category->name

Return DataFrame DataFrame filtered

mgkit.mappings.pandas_map.**make_stat_table**(*dataframes*, *roots*)

Produces a `pandas.DataFrame` that summarise the supplied DataFrames. The stats include mean, stdev and coefficient of variation for each root taxon.

Parameters

- **dataframes** (*iterable*) – iterable of DataFrame instances
- **roots** (*iterable*) – list of root taxa to which each table belongs

Return DataFrame returns a DataFrame instance

mgkit.mappings.taxon module

Module used to map taxon_id to different levels in the taxonomy.

mgkit.mappings.taxon.**map_taxon_by_id_list**(*taxon_id*, *map_ids*, *func*)

Maps a taxon_id to a list of taxon IDs, using the function supplied.

Parameters

- **taxon_id** (*int*[293]) – taxon ID to map
- **map_ids** (*iterable*) – list of taxon IDs to which the taxon_id will be mapped.
- **func** (*func*) – function used to map the IDs, accepts two taxon IDs

Results:

generator: generator expression of all IDs in map_ids to which taxon_id can be mapped.

Example

If mapping a taxon (Prevotella ruminicola) to Prevotella or Clostridium, using as *func* *mgkit.taxon.is_ancestor()* and taxonomy is an instance of *mgkit.taxon.UniprotTaxonomy*.

[290] https://docs.python.org/3/library/stdtypes.html#dict
[291] https://docs.python.org/3/library/stdtypes.html#str
[292] https://docs.python.org/3/library/stdtypes.html#dict
[293] https://docs.python.org/3/library/functions.html#int

```
>>> import functools
>>> from mgkit.taxon import is_ancestor
>>> func = functools.partial(is_ancestor, taxonomy)
>>> list(map_taxon_by_id_list(839, [838, 1485], func))
[838]
```

mgkit.mappings.utils module

Utilities to map genes

mgkit.mappings.utils.**count_genes_in_mapping**(*gene_lists*, *labels*, *mapping*, *normalise=False*)

Maps lists of ids to a mapping dictionary, returning a `pandas.DataFrame` in which the rows are the labels provided and the columns the categories to which the ids map. Each element of the matrix label-category is the sum of all ids in the relative gene list that maps to the specific category.

Parameters

- **gene_lists** (*iterable*) – an iterable in which each element is a iterable of ids that can be mapped to mapping
- **labels** (*iterable*) – an iterable of strings that defines the labels to be used in the resulting rows in the `pandas.DataFrame`; must have the same length as gene_lists
- **mapping** (*dict*[294]) – a dictionary in the form: gene_id->[cat1, cat2, .., catN]
- **normalise** (*bool*[295]) – if True the counts are normalised over the total for each row.

Returns a `pandas.DataFrame` instance

mgkit.mappings.utils.**group_annotation_by_mapping**(*annotations*, *mapping*, *attr='ko'*)

Group annotations by mapping dictionary

Parameters

- **annotations** (*iterable*) – iterable of `gff.GFFKeg` instances
- **mapping** (*dict*[296]) – dictionary with mappings for the attribute requested
- **attr** (*str*[297]) – attribute of the annotation to be used as key in mapping

Return dict dictionary category->annotations

Module contents

mgkit.net package

Submodules

mgkit.net.embl module

Access EMBL Services

exception mgkit.net.embl.**EntryNotFound**

Bases: `exceptions.Exception`

Raised if at least one entry was not found by *get_sequences_by_ids()*. `NOT_FOUND` is used to check if any entry wasn't downloaded.

[294] https://docs.python.org/3/library/stdtypes.html#dict
[295] https://docs.python.org/3/library/functions.html#bool
[296] https://docs.python.org/3/library/stdtypes.html#dict
[297] https://docs.python.org/3/library/stdtypes.html#str

exception mgkit.net.embl.**NoEntryFound**
Bases: exceptions.Exception

Raised if no sequences where found by *get_sequences_by_ids()*, the check is based on the NONE_FOUND variable.

mgkit.net.embl.**datawarehouse_search**(*query*, *domain='sequence'*, *result='sequence_release'*, *display='fasta'*, *offset=0*, *length=100000*, *contact=None*, *download='gzip'*, *url='http://www.ebi.ac.uk/ena/data/warehouse/search?'*, *fields=None*)
Changed in version 0.2.3: added *fields* parameter to retrieve tab separated information

New in version 0.1.13.

Perform a datawarehouse search on EMBL dbs. Instructions on the query language used to query the datawarehouse are available at this page[298] with more details about the databases domains at this page[299]

Parameters

- **query** (*str*[300]) – query for the search enging
- **domain** (*str*[301]) – database domain to search
- **result** (*str*[302]) – domain result requested
- **display** (*str*[303]) – display option (format to retrieve the entries)
- **offset** (*int*[304]) – the offset of the search results, defaults to the first
- **length** (*int*[305]) – number of results to retrieve at the specified offset and the limit is automatically set a 100,000 records for query
- **contact** (*str*[306]) – email of the user
- **download** (*str*[307]) – type of response. Gzip responses are automatically decompressed
- **url** (*str*[308]) – base URL for the resource
- **fields** (*None, iterable*) – must be an iterable of fields to be returned if display is set to *report*

Returns the raw request

Return type str[309]

Examples

Querying EMBL for all sequences of type rRNA of the *Clostridium* genus. Only from the EMBL release database in fasta format:

```
>>> query = 'tax_tree(1485) AND mol_type="rRNA"'
>>> result = 'sequence_release'
>>> display = 'fasta'
```

(continues on next page)

[298] http://www.ebi.ac.uk/ena/about/browser#data_warehouse
[299] http://www.ebi.ac.uk/ena/data/warehouse/usage
[300] https://docs.python.org/3/library/stdtypes.html#str
[301] https://docs.python.org/3/library/stdtypes.html#str
[302] https://docs.python.org/3/library/stdtypes.html#str
[303] https://docs.python.org/3/library/stdtypes.html#str
[304] https://docs.python.org/3/library/functions.html#int
[305] https://docs.python.org/3/library/functions.html#int
[306] https://docs.python.org/3/library/stdtypes.html#str
[307] https://docs.python.org/3/library/stdtypes.html#str
[308] https://docs.python.org/3/library/stdtypes.html#str
[309] https://docs.python.org/3/library/stdtypes.html#str

(continued from previous page)

```
>>> data = embl.datawarehouse_search(query, result=result,
... display=display)
>>> len(data)
35919
```

Each entry taxon_id from the same data can be retrieved using *report* as the *display* option and *fields* an iterable of fields to just ('accession', tax_id'):

```
>>> query = 'tax_tree(1485) AND mol_type="rRNA"'
>>> result = 'sequence_release'
>>> display = 'report'
>>> fields = ('accession', 'tax_id')
>>> data = embl.datawarehouse_search(query, result=result,
    display=display, fields=fields)
```

mgkit.net.embl.**dbfetch**(*embl_ids*, *db='embl'*, *contact=None*, *out_format='seqxml'*, *num_req=10*)

New in version 0.1.12.

Function that allows to use dbfetch service (REST). More information on the output formats and the database available at the service page[310]

Parameters

- **embl_ids** (*str*[311], *iterable*) – list or single sequence id to retrieve
- **db** (*str*[312]) – database from which retrieve the sequence data
- **contact** (*str*[313]) – email contact to use as per EMBL guidlines
- **out_format** (*str*[314]) – output format, depends on database
- **num_req** (*int*[315]) – number of ids per request

Returns a list with the results from each request sent. Each request sent has a maximum number *num_req* of ids, so the number of items in the list depends by the number of ids in *embl_ids* and the value of *num_req*.

Return type list[316]

mgkit.net.embl.**get_sequences_by_ids**(*embl_ids*, *contact=None*, *out_format='fasta'*, *num_req=10*, *embl_db='embl_cds'*, *compress=True*, *strict=False*)

Downloads entries using EBI REST API. It can download one entry at a time or accept an iterable and all sequences will be downloaded in batches of at most num_req.

It's fairly general, so can be customised, from the DB used to the output format: all batches are simply concatenate.

Note: There are some checks on the some errors reported by the EMBL api, but not documented, in particular two errors, which are just reported as text lines in the fasta file (the only one tested at this time).

The are two possible cases:

- if no entry was found *NoEntryFound* will be raised.
- if at least one entry wasn't found:
  - if strict is False (the default) the error will be just logged as a debug message

[310] http://www.ebi.ac.uk/Tools/dbfetch/syntax.jsp
[311] https://docs.python.org/3/library/stdtypes.html#str
[312] https://docs.python.org/3/library/stdtypes.html#str
[313] https://docs.python.org/3/library/stdtypes.html#str
[314] https://docs.python.org/3/library/stdtypes.html#str
[315] https://docs.python.org/3/library/functions.html#int
[316] https://docs.python.org/3/library/stdtypes.html#list

– if strict is True *EntryNotFound* is raised

Parameters

- **embl_ids** (*iterable, str*[317]) – list of ids to download
- **contact** (*str*[318]) – email address to be passed in the query
- **format** (*str*[319]) – format of the entry
- **num_req** (*int*[320]) – number of entries to download with each request
- **embl_db** (*str*[321]) – db to which the ids refer to
- **compress** (*bool*[322]) – if True, the function tries to obtain a compressed response and decompress it on the fly
- **strict** (*bool*[323]) – if True, a check on the number of entries retrieved is performed

Returns the entries requested

Return type str[324]

Raises

- *EntryNotFound* – if at least an entry was not found
- *NoEntryFound* – if NO entry were found

Warning: The number of sequences that can be downloaded at a time is 11, it seems, since the returned sequences for each request was at most 11. I didn't find any mention of this in the API docs, but it may be a restriction that's temporary.

mgkit.net.pfam module

New in version 0.2.3.

This module defines routine to access Pfam information using a network connection

mgkit.net.pfam.**get_pfam_families**(*key='id'*)

New in version 0.2.3.

Gets a dictionary with the accession/id/description of Pfam families from Pfam. This list can be accessed using the URL: http://pfam.xfam.org/families?output=text

The output is a tab separated file where the fields are:

- ACCESSION
- ID
- DESCRIPTION

Parameters **key** (*str*[325]) – if the value is *id*, the key of the dictionary is the ID, otherwise ID swaps position with ACCESSION (the new key)

[317] https://docs.python.org/3/library/stdtypes.html#str
[318] https://docs.python.org/3/library/stdtypes.html#str
[319] https://docs.python.org/3/library/stdtypes.html#str
[320] https://docs.python.org/3/library/functions.html#int
[321] https://docs.python.org/3/library/stdtypes.html#str
[322] https://docs.python.org/3/library/functions.html#bool
[323] https://docs.python.org/3/library/functions.html#bool
[324] https://docs.python.org/3/library/stdtypes.html#str
[325] https://docs.python.org/3/library/stdtypes.html#str

Returns by default the function returns a dictionary that uses the ID as key, while the value is a tuple (ACCESSION, DESCRIPTION). ID is the default because the *hmmer2gff - Convert HMMER output to GFF* script output uses ID as *gene_id* value when using the HMM provided by Pfam

Return type dict[326]

mgkit.net.uniprot module

Contains function and constants for Uniprot access

mgkit.net.uniprot.**get_gene_info**(*gene_ids*, *columns*, *max_req=50*, *contact=None*)

New in version 0.1.12.

Get informations about a list of genes. it uses *query_uniprot()* to send the request and format the response in a dictionary.

Parameters

- **gene_ids** (*iterable, str*[327]) – gene id(s) to get informations for
- **columns** (*list*[328]) – list of columns
- **max_req** (*int*[329]) – number of maximum *gene_ids* per request
- **contact** (*str*[330]) – email address to be passed in the query (requested Uniprot API)

Returns dictionary where the keys are the *gene_ids* requested and the values are dictionaries with the names of the *columns* requested as keys and the corresponding values, which can be lists if the values are are semicolon separated strings.

Return type dict[331]

Example

To get the taxonomy ids for some genes:

```
>>> uniprot.get_gene_info(['Q09575', 'Q8DQI6'], ['organism-id'])
{'Q09575': {'organism-id': '6239'}, 'Q8DQI6': {'organism-id': '171101'}}
```

mgkit.net.uniprot.**get_ko_to_eggnog_mappings**(*ko_ids*, *contact=None*)

New in version 0.1.14.

It's not possible to map in one go KO IDs to eggNOG IDs via the API in Uniprot. This function uses *query_uniprot()* to get all Uniprot IDs requested and the return a dictionary with all their eggNOG IDs they map to.

Parameters

- **ko_ids** (*iterable*) – an iterable of KO IDs
- **contact** (*str*[332]) – email address to be passed in the query (requested Uniprot API)

Returns The format of the resulting dictionary is ko_id -> {eggnog_id1, ..}

Return type dict[333]

[326] https://docs.python.org/3/library/stdtypes.html#dict
[327] https://docs.python.org/3/library/stdtypes.html#str
[328] https://docs.python.org/3/library/stdtypes.html#list
[329] https://docs.python.org/3/library/functions.html#int
[330] https://docs.python.org/3/library/stdtypes.html#str
[331] https://docs.python.org/3/library/stdtypes.html#dict
[332] https://docs.python.org/3/library/stdtypes.html#str
[333] https://docs.python.org/3/library/stdtypes.html#dict

mgkit.net.uniprot.**get_mappings**(*entry_ids*, *db_from='ID'*, *db_to='EMBL'*, *out_format='tab'*, *contact=None*)

Gets mapping of genes using Uniprot REST API. The db_from and db_to values are the ones accepted by Uniprot API. The same applies to out_format, the only processed formats are 'list', which returns a list of the mappings (should be used with one gene only) and 'tab', which returns a dictionary with the mapping. All other values returns a string with the newline stripped.

Parameters

- **entry_ids** (*iterable*) – iterable of ids to be mapped (there's a limit) to the maximum length of a HTTP request, so it should be less than 50
- **db_from** (*str*[334]) – string that identify the DB for elements in entry_ids
- **db_to** (*str*[335]) – string that identify the DB to which map entry_ids
- **out_format** (*str*[336]) – format of the mapping; 'list' and 'tab' are processed
- **contact** (*str*[337]) – email address to be passed in the query (requested Uniprot API)

Returns tuple, dict or str depending on out_format value

mgkit.net.uniprot.**get_sequences_by_ko**(*ko_id*, *taxonomy*, *contact=None*, *reviewed=True*)

Gets sequences from Uniprot, restricting to the taxon id passed.

Parameters

- **ko_id** (*str*[338]) – KO id of the sequences to download
- **taxonomy** (*int*[339]) – id of the taxon
- **contact** (*str*[340]) – email address to be passed in the query (requested by Uniprot API)
- **reviewed** (*bool*[341]) – if the sequences requested must be reviewed

Returns string with the fasta file downloaded

mgkit.net.uniprot.**get_uniprot_ec_mappings**(*gene_ids*, *contact=None*)

New in version 0.1.14.

Shortcut to download EC mapping of Uniprot IDs. Uses *get_gene_info()* passing the correct column (*ec*).

mgkit.net.uniprot.**ko_to_mapping**(*ko_id*, *query*, *columns*, *contact=None*)

Returns the mappings to the supplied KO. Can be used for any id, the query format is free as well as the columns returned. The only restriction is using a tab format, that is parsed.

Parameters

- **ko_id** (*str*[342]) – id used in the query
- **query** (*str*[343]) – query passed to the Uniprot API, ko_id is replaced using `str.format()`
- **column** (*str*[344]) – column used in the results table used to map the ids
- **contact** (*str*[345]) – email address to be passed in the query (requested Uniprot API)

[334] https://docs.python.org/3/library/stdtypes.html#str
[335] https://docs.python.org/3/library/stdtypes.html#str
[336] https://docs.python.org/3/library/stdtypes.html#str
[337] https://docs.python.org/3/library/stdtypes.html#str
[338] https://docs.python.org/3/library/stdtypes.html#str
[339] https://docs.python.org/3/library/functions.html#int
[340] https://docs.python.org/3/library/stdtypes.html#str
[341] https://docs.python.org/3/library/functions.html#bool
[342] https://docs.python.org/3/library/stdtypes.html#str
[343] https://docs.python.org/3/library/stdtypes.html#str
[344] https://docs.python.org/3/library/stdtypes.html#str
[345] https://docs.python.org/3/library/stdtypes.html#str

Note: each mapping in the column is separated by a ;

mgkit.net.uniprot.**parse_uniprot_response**(*data*, *simple=True*)
New in version 0.1.12.

Parses raw response from a Uniprot query (tab format only) from functions like *query_uniprot()* into a dictionary. It requires that the first column is the entry id (or any other unique id).

Parameters

- **data** (*str*[346]) – string response from Uniprot
- **simple** (*bool*[347]) – if True and the number of columns is 1, the dictionary returned has a simplified structure

Returns The format of the resulting dictionary is entry_id -> {column1 -> value, column2 -> value, ..} unless there's only one column and *simple* is True, in which case the value is equal to the value of the only column.

Return type dict[348]

mgkit.net.uniprot.**query_uniprot**(*query*, *columns=None*, *format='tab'*, *limit=None*, *contact=None*, *baseurl='http://www.uniprot.org/uniprot/'*)
New in version 0.1.12.

Changed in version 0.1.13: added *baseurl* and made *columns* a default argument

Queries Uniprot, returning the raw response in tbe format specified. More informations at the page[349]

Parameters

- **query** (*str*[350]) – query to submit, as put in the input box
- **columns** (*None, iterable*) – list of columns to return
- **format** (*str*[351]) – response format
- **limit** (*int*[352], *None*) – number of entries to return or *None* to request all entries
- **contact** (*str*[353]) – email address to be passed in the query (requested Uniprot API)
- **baseurl** (*str*[354]) – base url for the REST API, can be either `UNIPROT_GET` or `UNIPROT_TAXONOMY`

Returns raw response from the query

Return type str[355]

Example

To get the taxonomy ids for some genes:

```
>>> uniprot.query_uniprot('Q09575 OR Q8DQI6', ['id', 'organism-id'])
'Entry\tOrganism ID\nQ8DQI6\t171101\nQ09575\t6239\n'
```

[346] https://docs.python.org/3/library/stdtypes.html#str
[347] https://docs.python.org/3/library/functions.html#bool
[348] https://docs.python.org/3/library/stdtypes.html#dict
[349] http://www.uniprot.org/faq/28
[350] https://docs.python.org/3/library/stdtypes.html#str
[351] https://docs.python.org/3/library/stdtypes.html#str
[352] https://docs.python.org/3/library/functions.html#int
[353] https://docs.python.org/3/library/stdtypes.html#str
[354] https://docs.python.org/3/library/stdtypes.html#str
[355] https://docs.python.org/3/library/stdtypes.html#str

> **Warning:** because of limits in the length of URLs, it's advised to limit the length of the query string.

mgkit.net.utils module

Utility functions for the network package

mgkit.net.utils.**url_open**(*url*, *data=None*, *compress=True*, *agent=None*)

Utility function that compresses the request and add the user agent header if supplied.

Parameters

- **url** (*str*[356]) – url to request
- **data** (*str*[357]) – data to add to the request
- **compress** (*bool*[358]) – if the response should be compressed
- **agent** (*str*[359]) – if supplied, the 'User-Agent' header we'll be added to the request

Returns the response handle

mgkit.net.utils.**url_read**(*url*, *data=None*, *compress=True*, *agent=None*)

Utility function that compresses the request and add the user agent header if supplied.

Wrapper of *url_open()* which reads the full response

Parameters

- **url** (*str*[360]) – url to request
- **data** (*str*[361]) – data to add to the request
- **compress** (*bool*[362]) – if the response should be compressed
- **agent** (*str*[363]) – if supplied, the 'User-Agent' header we'll be added to the request

Returns the response data

Module contents

Package with functions/classes used in accessing network resources

mgkit.plots package

Submodules

mgkit.plots.abund module

New in version 0.1.15.

Module to plot relative abundances in a 1D or 3D projection

mgkit.plots.abund.**col_func_firstel**(*key*, *colors=None*)

mgkit.plots.abund.**col_func_name**(*key*, *func=None*, *colors=None*)

[356] https://docs.python.org/3/library/stdtypes.html#str
[357] https://docs.python.org/3/library/stdtypes.html#str
[358] https://docs.python.org/3/library/functions.html#bool
[359] https://docs.python.org/3/library/stdtypes.html#str
[360] https://docs.python.org/3/library/stdtypes.html#str
[361] https://docs.python.org/3/library/stdtypes.html#str
[362] https://docs.python.org/3/library/functions.html#bool
[363] https://docs.python.org/3/library/stdtypes.html#str

mgkit.plots.abund.**col_func_taxon**(*taxon_id*, *taxonomy*, *anc_ids*, *colpal*)

mgkit.plots.abund.**draw_1d_grid**(*ax, labels=['LAM', 'SAM'], fontsize=22*)

Changed in version 0.2.0: reworked internals and changed defaults

Draws a 1D axis, to display propotions.

Parameters

- **ax** – an axis instance
- **labels** (*iterable*) – list of string to be put for the axes
- **fontsize** (*float*[364]) – font size for the labels, the tick font size is equal to *0.75 * fontsize*

mgkit.plots.abund.**draw_axis_internal_triangle**(*ax*, *color='r'*, *linewidth=2.0*)

New in version 0.2.5.

Draws a triangle that indicates the 50% limit for all 3 samples

Parameters

- **ax** – axis to use
- **color** (*str*[365], *float*[366], *tuple*[367]) – color used to draw the triangle
- **linewidth** (*float*[368]) – line width

mgkit.plots.abund.**draw_circles**(*ax*, *data*, *col_func=<function col_func_name>*, *csize=200*, *alpha=0.5*, *sizescale=None*, *order=None*, *linewidths=0.0*, *edgecolor='none'*)

Changed in version 0.2.0: changed internals and added return value

Draws a scatter plot over either a planar-simplex projection, if the number of coordinates is 3, or in a 1D axis.

If the number of coordinates is 3, *project_point()* is used to project the point in 2 coordinates. The coordinates are converted in proportions internally.

Parameters

- **ax** – axis to plot on
- **data** (*pandas.DataFrame*) – a DataFrame with 2 for a 1D plot or 3 columns for a planar-simplex
- **col_func** (*func*) – a function that accept a parameter, an element of the DataFrame index and returns a colour for it
- **csize** (*int*[369]) – the base size of the circles
- **alpha** (*float*[370]) – transparency of the circles, between 0 and 1 included
- **sizescale** (*None, pandas.Series*) – a Series or dictionary with the same elements as the Index of *data*, whose values are the size factors that are multiplied to *csize*. If **None**, the size of the circles is equal to *csize*
- **order** (*None, iterable*) – iterable with the elements of *data* Index, to specify the order in which the circles must be plotted. If None, the order is the same as *data.index*
- **linewidths** (*float*[371]) – width of the circle line

[364] https://docs.python.org/3/library/functions.html#float
[365] https://docs.python.org/3/library/stdtypes.html#str
[366] https://docs.python.org/3/library/functions.html#float
[367] https://docs.python.org/3/library/stdtypes.html#tuple
[368] https://docs.python.org/3/library/functions.html#float
[369] https://docs.python.org/3/library/functions.html#int
[370] https://docs.python.org/3/library/functions.html#float
[371] https://docs.python.org/3/library/functions.html#float

- **edgecolor** (*str*[372]) – color of the circle line

Returns the return value of matplotlib *scatter*

Return type PathCollection

Note: To **not** have circle lines, *edgecolor* must be *'none'* and *linewidths* equal *0*

mgkit.plots.abund.**draw_triangle_grid**(*ax, labels=['LAM', 'SAM', 'EAM'], linewidth=1.0, styles=['-', ':', '–'], fontsize=22*)

Changed in version 0.2.0: reworked internals and changed defaults

Draws a triangle as axes, for a planar-simplex projection.

Parameters

- **ax** – an axis instance
- **labels** (*iterable*) – list of string to be put for the axes
- **styles** (*None, iterable*) – either *None* for solid lines or matplotlib line markers. These are in sync between the internal lines and the axes.
- **linewidth** (*float*[373]) – line width for the axes, the internal lines are equal to *0.75 * linewidth*
- **fontsize** (*float*[374]) – font size for the labels, the tick font size is equal to *0.75 * fontsize*

mgkit.plots.abund.**project_point**(*point*)

Project a tuple containing coordinates (i.e. x, y, z) to planar-simplex.

Parameters **point** (*tuple*[375]) – contains the three coordinates to project

Returns the projected point in a planar-simplex

Return type tuple[376]

mgkit.plots.boxplot module

New in version 0.1.14.

Code related to boxplots

mgkit.plots.boxplot.**add_values_to_boxplot**(*dataframe, ax, plot_data, plot_order, data_colours=None, alpha=0.5, s=80, marker='o', linewidth=0.01, box_vert=False*)

New in version 0.1.13.

Changed in version 0.1.14: added *box_vert* parameter

Changed in version 0.1.16: changed default value for *linewidth*

Adds the values of a dataframe used in *boxplot_dataframe()* to the plot. *linewidth* must be higher than 0 if a marker like | is used.

A list of markers is available at this page[377]

[372] https://docs.python.org/3/library/stdtypes.html#str
[373] https://docs.python.org/3/library/functions.html#float
[374] https://docs.python.org/3/library/functions.html#float
[375] https://docs.python.org/3/library/stdtypes.html#tuple
[376] https://docs.python.org/3/library/stdtypes.html#tuple
[377] http://matplotlib.org/api/markers_api.html

Warning: Contrary to *boxplot_dataframe()*, the boxplot default is horizontal (*box_vert*). The default will change in a later version.

Parameters

- **dataframe** – dataframe with the values to plot
- **ax** – an axis instance
- **plot_data** – return value from *boxplot_dataframe()*
- **plot_order** (*iterable*) – row order used to plot the boxes
- **data_colours** (*dict*[378]) – colors used for the values
- **alpha** (*float*[379]) – alpha value for the colour
- **s** (*int*[380]) – size of the marker drawn
- **marker** (*str*[381]) – one of the accepted matplotlib markers
- **linewidth** (*float*[382]) – width of the line used to draw the marker
- **box_vert** (*bool*[383]) – specify if the original boxplot is vertical or not

mgkit.plots.boxplot.**add_significance_to_boxplot**(*sign_indices*, *ax*, *pos*, *box_vert=True*, *fontsize=16*)

New in version 0.1.16.

Add significance groups to boxplots

Parameters

- **sign_indices** (*iterable*) – iterable in which each element is a tuple; each element of the tuple is the numerical index of the position of the significant boxplot
- **ax** – an axis instance
- **pos** (*tuple*[384]) – the 2 values are the coordinates for the top line, and the the lowest bound for the whisker
- **box_vert** (*bool*[385]) – if the boxplot is vertical
- **fontsize** (*float*[386]) – size for the * (star)

mgkit.plots.boxplot.**boxplot_dataframe_multindex**(*dataframe*, *axes*, *plot_order=None*, *label_map=None*, *fonts=None*, *fill_box=True*, *colours=None*, *data_colours=None*, *box_vert=True*)

New in version 0.1.13.

Todo: documentation

[378] https://docs.python.org/3/library/stdtypes.html#dict
[379] https://docs.python.org/3/library/functions.html#float
[380] https://docs.python.org/3/library/functions.html#int
[381] https://docs.python.org/3/library/stdtypes.html#str
[382] https://docs.python.org/3/library/functions.html#float
[383] https://docs.python.org/3/library/functions.html#bool
[384] https://docs.python.org/3/library/stdtypes.html#tuple
[385] https://docs.python.org/3/library/functions.html#bool
[386] https://docs.python.org/3/library/functions.html#float

The function draws a series of boxplots from a DataFrame object, whose order is directed by the iterable plot_order. The columns of each DataFrame row contains the values for each boxplot. An axes object is needed.

Parameters

- **dataframe** – dataframe to plot
- **plot_order** (*iterable*) – row order used to plot the boxes
- **axes** – an axes instance
- **label_map** (*dict*[387]) – a map that converts the items in plot_order to a label used on the plot X axes
- **fonts** (*dict*[388]) – dictionary with properties for x axis labels, DEFAULT_BOXPLOT_FONTCONF is used by default
- **fill_box** (*bool*[389]) – if True each box is filled with the same colour of its outline
- **colours** (*dict*[390]) – dictionary with properties for each boxplot if data_colours is None, whi overrides box, whiskers and fliers. Defaults to DEFAULT_BOXPLOT_COLOURS
- **data_colours** (*dict*[391]) – dictionary of colours for each boxplot, a set of colours can be obtained using func:*map_taxon_to_colours*

Returns the plot data same as matplotlib boxplot function

mgkit.plots.boxplot.**boxplot_dataframe**(*dataframe*, *plot_order*, *ax*, *label_map=None*, *fonts=None*, *fill_box=True*, *colours=None*, *data_colours=None*, *box_vert=True*, *widths=0.5*)

New in version 0.1.7: To move from an all-in-one drawing to a more modular one.

Changed in version 0.1.13: added box_vert parameter

Changed in version 0.1.16: added *widths* parameter

The function draws a series of boxplots from a DataFrame object, whose order is directed by the iterable plot_order. The columns of each DataFrame row contains the values for each boxplot. An ax object is needed.

Parameters

- **dataframe** – dataframe to plot
- **plot_order** (*iterable*) – row order used to plot the boxes
- **ax** – an axis instance
- **label_map** (*dict*[392]) – a map that converts the items in plot_order to a label used on the plot X ax
- **fonts** (*dict*[393]) – dictionary with properties for x axis labels, DEFAULT_BOXPLOT_FONTCONF is used by default
- **fill_box** (*bool*[394]) – if True each box is filled with the same colour of its outline

[387] https://docs.python.org/3/library/stdtypes.html#dict
[388] https://docs.python.org/3/library/stdtypes.html#dict
[389] https://docs.python.org/3/library/functions.html#bool
[390] https://docs.python.org/3/library/stdtypes.html#dict
[391] https://docs.python.org/3/library/stdtypes.html#dict
[392] https://docs.python.org/3/library/stdtypes.html#dict
[393] https://docs.python.org/3/library/stdtypes.html#dict
[394] https://docs.python.org/3/library/functions.html#bool

- **colours** (*dict*[395]) – dictionary with properties for each boxplot if data_colours is None, whi overrides box, whiskers and fliers. Defaults to DEFAULT_BOXPLOT_COLOURS
- **data_colours** (*dict*[396]) – dictionary of colours for each boxplot, a set of colours can be obtained using func:*map_taxon_to_colours*
- **box_vert** (*bool*[397]) – if False the boxplots are drawn horizontally
- **widths** (*float*[398]) – width (scalar or array) of the boxplots width(s)

Returns the plot data; same as matplotlib boxplot function

mgkit.plots.colors module

New in version 0.1.14.

Contains code related to colour

mgkit.plots.colors.**float_to_hex_color**(*r*, *g*, *b*)
New in version 0.1.14.

Converts RGB float values to Hexadecimal value string

mgkit.plots.colors.**palette_float_to_hex**(*palette*)
New in version 0.1.16.

Applies *float_to_hex_color()* to an iterable of colors

mgkit.plots.heatmap module

New in version 0.1.14.

Code related to heatmaps.

mgkit.plots.heatmap.**baseheatmap**(*data*, *ax*, *norm=None*, *cmap=None*, *xticks=None*, *yticks=None*, *fontsize=18*, *meshopts=None*, *annot=False*, *annotopts=None*)
Changed in version 0.2.3: added *annot* and *annot_args* arguments

A basic heatmap using `matplotlib.pyplot.pcolormesh()`. It expect a `pandas.DataFrame`.

Note: Rows a plot bottom to up, while the columns left to right. Change the order of the DataFrame if needed.

Parameters

- **data** (*pandas.DataFrame*) – matrix to plot. The DataFrame labels are used
- **ax** – axes to use
- **norm** – if needed, `matplotlib.colors.BoundaryNorm` or `matplotlib.colors.Normalize` can be used to fine tune the colors
- **cmap** (*None*, *matplotlib.colors.ListedColormap*) – color map to use
- **xticks** (*None*, *dict*[399]) – dictionary with additional options to pass to *set_xticklabels*

[395] https://docs.python.org/3/library/stdtypes.html#dict
[396] https://docs.python.org/3/library/stdtypes.html#dict
[397] https://docs.python.org/3/library/functions.html#bool
[398] https://docs.python.org/3/library/functions.html#float
[399] https://docs.python.org/3/library/stdtypes.html#dict

- **yticks** (*None*, *dict*[400]) – dictionary with additional options to pass to *set_yticklabels*
- **fontsize** (*int*[401]) – font size to use for the labels
- **meshopts** (*None*, *dict*[402]) – additional options to pass to `matplotlib.pyplot.pcolormesh()`
- **annot** (*bool*[403]) – if True the values of the matrix will be added
- **annot_args** (*None*, *dict*[404]) – dictionary with the options for the annotations. The option *format* is a function that returns the formatted number, defaults to a number with no decimal part

Returns the return value of `matplotlib.pyplot.pcolormesh()`

Return type matplotlib.collections.QuadMesh

`mgkit.plots.heatmap.`**grouped_spine**(*groups*, *labels*, *ax*, *which='y'*, *spine='right'*, *spine_opts=None*, *start=0*)

Changed in version 0.2.0: added *va*, *ha* keys to *spine_opts*, changed the label positioning

Changed in version 0.2.5: added *start* parameter

Changes the spine of an heatmap axis given the groups of labels.

Note: It should work for any plot, but was not tested

Parameters

- **groups** (*iterable*) – a nested list where each is element is a list containing the labels belong to that group.
- **labels** (*iterable*) – an iterable with the labels of the groups. Needs to be in the same order as groups
- **ax** – axis to use (same as heatmap)
- **which** (*str*[405]) – to specify the axis, either *x* or *y*
- **spine** (*str*[406]) – position of the spine. if *which* is **x** accepted values are *top* and *bottom*, if which is **y** *left* and *right* are accepted
- **spine_opts** (*dict*[407]) – additional options to pass to the spine class
- **start** (*int*[408]) – the start coordinate for the grouped spine. Defaults to 0

`mgkit.plots.heatmap.`**dendrogram**(*data*, *ax*, *method='complete'*, *orientation='top'*, *use_dist=True*, *dist_func=<function pdist>*)

Changed in version 0.1.16: added *use_dist* and *dist_func* parameters

Plots a dendrogram of the clustered rows of the given matrix; if the columns are to be clustered, the transposed matrix needs to be passed.

Parameters

- **data** (*pandas.DataFrame*) – matrix to plot. The DataFrame labels are used

[400] https://docs.python.org/3/library/stdtypes.html#dict
[401] https://docs.python.org/3/library/functions.html#int
[402] https://docs.python.org/3/library/stdtypes.html#dict
[403] https://docs.python.org/3/library/functions.html#bool
[404] https://docs.python.org/3/library/stdtypes.html#dict
[405] https://docs.python.org/3/library/stdtypes.html#str
[406] https://docs.python.org/3/library/stdtypes.html#str
[407] https://docs.python.org/3/library/stdtypes.html#dict
[408] https://docs.python.org/3/library/functions.html#int

- **ax** – axes to use
- **method** (*str*[409]) – clustering method used, internally `scipy.cluster.hierarchy.linkage()` is used.
- **orientation** (*str*[410]) – direction for the plot. *top*, *bottom*, *left* and *right* are accepted; *top* will draw the leaves at the bottom.
- **use_dist** (*bool*[411]) – if True, the function *dist_func* will be applied to *data* to get a distance matrix
- **dist_func** (*func*) – distance function to be used

Returns The dendrogram plotted, as returned by `scipy.cluster.hierarchy.dendrogram()`

`mgkit.plots.heatmap.`**heatmap_clustered**(*data*, *figsize=(10, 5)*, *cmap=None*, *norm=None*)

Plots a heatmap clustered on both rows and columns.

Parameters

- **data** (*pandas.DataFrame*) – matrix to plot. The DataFrame labels are used
- **figsize** (*tuple*[412]) – passed to *mgkit.plots.utils.get_grid_figure()*
- **cmap** (*None, matplotlib.colors.ListedColormap*) – color map to use
- **norm** – if needed, `matplotlib.colors.BoundaryNorm` or `matplotlib.colors.Normalize` can be used to fine tune the colors

mgkit.plots.unused module

New in version 0.1.14.

Code deprecated or untested

`mgkit.plots.unused.`**barchart_categories**(*data*, *colours=None*, *title=''*, *tickfont='small'*, *xlabel_dict=None*, *barlabel_dict=None*, *width=0.9*, *rotation='vertical'*, *file_name=None*, *fig_size=None*, *fig_aspect=None*)

Parameters

- **data** – DataFrame where the number of rows indicates how many bars will plotted per column
- **colours** – must be equal the number of data rows if supplied or it will be blue by default
- **title** – chart title
- **tickfont** – font size for ticks (only for column axis)
- **xlabel_dict** – a mapping to the acual labels to use for the columns. Defaults to columns' names
- **barlabel_dict** – a mapping to the acual labels to use for the row. Defaults to rows' names
- **width** – bar width

Returns axis instance

[409] https://docs.python.org/3/library/stdtypes.html#str
[410] https://docs.python.org/3/library/stdtypes.html#str
[411] https://docs.python.org/3/library/functions.html#bool
[412] https://docs.python.org/3/library/stdtypes.html#tuple

mgkit.plots.unused.**get_taxon_colors_new**(*taxa*, *taxonomy*, *default_colour='#ffff33'*)
Returns a dictionary of taxa and their assigned colours based on TAXON_COLOURS and the taxonomy provided. Uses the taxon.UniprotTaxonomy.get_taxon_root() to determine the root of a taxon.

Parameters

- **taxa** (*iterable*) – iterable of taxon ids
- **taxonomy** (UniprotTaxonomy) – taxonomy instance
- **default_colour** – colour used in case there's no known root for the taxon

Return dict dictionary mapping taxon_id to colour

mgkit.plots.unused.**lineplot_values_on_second_axis**(**args*, ***kwargs*)
Deprecated since version 0.1.13.

Adds a lineplot on a second axis using twinx

mgkit.plots.unused.**map_taxon_to_colours**(*taxa*, *taxonomy*, *default_colour='#ffff33'*)
Returns a dictionary of taxa and their assigned colours based on TAXON_COLOURS and the taxonomy provided. Uses the taxon.UniprotTaxonomy.get_taxon_root() to determine the root of a taxon.

Parameters

- **taxa** (*iterable*) – iterable of taxon ids
- **taxonomy** (UniprotTaxonomy) – taxonomy instance
- **default_colour** – colour used in case there's no known root for the taxon

Return dict dictionary mapping taxon_id to colour

mgkit.plots.unused.**plot_contig_assignment_bar**(*series*, *taxon_colours=None*, *log_scale=False*, *index=None*, *file_name=None*, *fig_aspect=None*, *xlabels_size=8*)
Plots barchart for contig assignment

Parameters

- **series** – pandas.Series instance with the data
- **taxon_colours** (*dict*[413]) – colour of the bars for each taxon
- **log_scale** (*bool*[414]) – if True the y axis is log scaled
- **fig_aspect** (*tuple*[415]) – tuple with figure size
- **xlabels_size** (*int*[416]) – size of the taxon labels
- **index** – optional pandas.Index used to reindex the series
- **file_name** (*str*[417]) – name of the file to write the graph to

mgkit.plots.unused.**plot_scatter_2d**(*data*, *labels*, *colours=None*, *pointsize=10.0*, *title=''*, *xlabel=''*, *ylabel=''*, *centers=None*, *marker='*'*, *marker_colour=None*, *markersize=None*, *hull_points=True*, *linewidth=0.2*, *legend=True*, *anno_center=True*)
Scatter plot in 2d. Used for cluster results

Parameters

- **data** (*array*) – numpy.array with shape n, 2
- **labels** (*array*) – labels to categorise samples

[413] https://docs.python.org/3/library/stdtypes.html#dict
[414] https://docs.python.org/3/library/functions.html#bool
[415] https://docs.python.org/3/library/stdtypes.html#tuple
[416] https://docs.python.org/3/library/functions.html#int
[417] https://docs.python.org/3/library/stdtypes.html#str

- **colours** (*dict*[418]) – dictionary whose keys are the labels and the values are valid `matplotlib` colours
- **pointsize** (*int*[419]) – point size
- **title** (*str*[420]) – plot title
- **xlabel** (*str*[421]) – label for x axis
- **ylabel** (*str*[422]) – label for y axis

Returns axis instance

`mgkit.plots.unused.`**plot_scatter_3d**(*data*, *labels*, *colours=None*, *pointsize=10.0*, *title=''*, *xlabel=''*, *ylabel=''*, *zlabel=''*)

Scatter plot in 3d. Used for cluster results

Parameters

- **data** (*array*) – `numpy.array` with shape n, 3
- **labels** (*array*) – labels to categorise samples
- **colours** (*dict*[423]) – dictionary whose keys are the labels and the values are valid `matplotlib` colours
- **pointsize** (*int*[424]) – point size
- **title** (*str*[425]) – plot title
- **xlabel** (*str*[426]) – label for x axis
- **ylabel** (*str*[427]) – label for y axis
- **zlabel** (*str*[428]) – label for z axis

Returns axis instance

`mgkit.plots.unused.`**scatter_gene_values**(*gene_dict*, *xlabel='Profile pN/pS'*, *ylabel='Rumen pN/pS'*, *title=''*, *colours=None*, *file_name=None*, *plot_order=None*, *line_colour='r'*, *max_limit=None*, *axes=None*)

Plots gene-taxon pN/pS from profiles against their observed values.

Parameters

- **gene_dict** (*dict*[429]) – dictionary that contains the data
- **xlabel** (*str*[430]) – label for x axis
- **ylabel** (*str*[431]) – label for y axis
- **title** (*str*[432]) – graph title
- **colours** – colours used in for the different datasets; defaults to TAXON_COLOURS

[418] https://docs.python.org/3/library/stdtypes.html#dict
[419] https://docs.python.org/3/library/functions.html#int
[420] https://docs.python.org/3/library/stdtypes.html#str
[421] https://docs.python.org/3/library/stdtypes.html#str
[422] https://docs.python.org/3/library/stdtypes.html#str
[423] https://docs.python.org/3/library/stdtypes.html#dict
[424] https://docs.python.org/3/library/functions.html#int
[425] https://docs.python.org/3/library/stdtypes.html#str
[426] https://docs.python.org/3/library/stdtypes.html#str
[427] https://docs.python.org/3/library/stdtypes.html#str
[428] https://docs.python.org/3/library/stdtypes.html#str
[429] https://docs.python.org/3/library/stdtypes.html#dict
[430] https://docs.python.org/3/library/stdtypes.html#str
[431] https://docs.python.org/3/library/stdtypes.html#str
[432] https://docs.python.org/3/library/stdtypes.html#str

- **file_name** (*str*[433]) – path to which the graph is to be saved (by default) it doesn't write to disk
- **plot_order** (*iterable*) – the order used in plotting the data points; default to the order of the gene_dict dictionary keys
- **coloUr** – valid colour for the lines in the plot
- **max_limit** (*float*[434]) – used to put a limit on the plot
- **axes** – optional axes used to draw the scatter plot

Returns the axis object used for the plot

mgkit.plots.utils module

New in version 0.1.14.

Misc code

mgkit.plots.utils.**get_grid_figure**(*rows*, *cols*, *dpi=300*, *figsize=(10, 20)*, ***kwd*)

New in version 0.1.13.

Simple wrapper to init a GridSpec figure

Parameters

- **rows** (*int*[435]) – number of rows
- **columns** (*int*[436]) – number of columns
- **dpi** (*int*[437]) – dpi used for the figure
- **figsize** (*tuple*[438]) – size of the figure in inches

Returns the figure and axes objects

Return type tuple[439]

mgkit.plots.utils.**get_single_figure**(*dpi=300*, *figsize=(10, 20)*, *aspect='auto'*)

Changed in version 0.1.14: added *aspect* parameter

Simple wrapper to init a single figure

Parameters

- **dpi** (*int*[440]) – dpi used for the figure
- **figsize** (*tuple*[441]) – size of the figure in inches
- **aspect** (*str*[442], *float*[443]) – aspect ratio to be passed to figure.add_subplot

Returns the figure and axes objects

Return type tuple[444]

[433] https://docs.python.org/3/library/stdtypes.html#str
[434] https://docs.python.org/3/library/functions.html#float
[435] https://docs.python.org/3/library/functions.html#int
[436] https://docs.python.org/3/library/functions.html#int
[437] https://docs.python.org/3/library/functions.html#int
[438] https://docs.python.org/3/library/stdtypes.html#tuple
[439] https://docs.python.org/3/library/stdtypes.html#tuple
[440] https://docs.python.org/3/library/functions.html#int
[441] https://docs.python.org/3/library/stdtypes.html#tuple
[442] https://docs.python.org/3/library/stdtypes.html#str
[443] https://docs.python.org/3/library/functions.html#float
[444] https://docs.python.org/3/library/stdtypes.html#tuple

mgkit.plots.utils.**legend_patches**(*labels*, *colors*)

New in version 0.3.1.

Makes handles (using matplotlib Patch) that can be passed to the legend method of a matplotlib axes instance

Parameters

- **labels** (*iterable*) – iterable that yields a label
- **colors** (*iterable*) – iterable that yields a valid matplotlib color

Returns list of patches that can be passed to the *handles* parameter in the *ax.legend* method

Return type list[445]

Module contents

New in version 0.1.14.

mgkit.snps package

Submodules

mgkit.snps.classes module

Manage SNP data.

class mgkit.snps.classes.**GeneSNP**(*gene_id=''*, *taxon_id=0*, *exp_syn=0*, *exp_nonsyn=0*, *coverage=None*, *snps=None*, *uid=None*, *json_data=None*)

Bases: *mgkit.snps.classes.RatioMixIn*

New in version 0.1.13.

Class defining gene and synonymous/non-synonymous SNPs.

It defines background synonymous/non-synonymous attributes and only has a method right now, which calculate pN/pS ratio. The method is added through a mixin object, so the ratio can be customised and be shared with the old implementation.

uid
str – unique id for the isoform (to be referenced in a GFF file)

gene_id
str – gene id

taxon_id
int – gene taxon

exp_syn
int – expected synonymous changes

exp_nonsyn
int – expected non-synonymous changes

coverage
int – gene coverage

snps
list – list of SNPs associated with the gene, each element is a tuple with the position (relative to the gene start), the second is the nucleotidic change and the third is the aa SNP type as defined by *SNPType*.

[445] https://docs.python.org/3/library/stdtypes.html#list

Note: The main difference with the `GeneSyn` is that all snps are kept and *syn* and *nonsyn* are not attributes but properties that return the count of synonymous and non-synonymous SNPs in the *snps* list.

Warning: This class uses more memory than `GeneSyn` because it doesn't use __slots__, it may be changed in later versions.

add(*other*)
Inplace addition of another instance values. No check for them being the same gene/taxon, it's up to the user to check that they can be added together.

Parameters **other** – instance of `GeneSyn` to add

add_snp(*position*, *change*, *snp_type=<SNPType.unknown: 0>*)
Adds a SNP to the list

Parameters

- **position** (*int*[446]) – SNP position, relative to the gene start
- **change** (*str*[447]) – nucleotidic change
- **snp_type** (*enum*) – one of the values defined in *SNPType*

coverage = None

exp_nonsyn = None

exp_syn = None

from_json(*data*)
Instantiate the instance with values from a json definition

Parameters **data** (*str*[448]) – json representation, as returned by *GeneSNP.to_json()*

gene_id = None

nonsyn
Returns the expected non-synonymous changes

snps = None

syn
Returns the expected synonymous changes

taxon_id = None

to_json()
Returns a json definition of the instance

Returns json representation of the instance

Return type str[449]

uid = None

class mgkit.snps.classes.**RatioMixIn**
Bases: object[450]

calc_ratio(*haplotypes=False*)
Changed in version 0.2.2: split the function to handle *flag_value* in another method

[446] https://docs.python.org/3/library/functions.html#int
[447] https://docs.python.org/3/library/stdtypes.html#str
[448] https://docs.python.org/3/library/stdtypes.html#str
[449] https://docs.python.org/3/library/stdtypes.html#str
[450] https://docs.python.org/3/library/functions.html#object

Calculate $\frac{pN}{pS}$ for the gene.

$$\frac{pN}{pS} = \frac{{}^{oN}/_{eN}}{{}^{oS}/_{eS}} \tag{7.3}$$

WHere:

- oN (number of non-synonymous - **nonsyn**)
- eN (expected number of non-synonymous - **exp_nonsyn**)
- oS (number of synonymous - **syn**)
- eS (expected number of synonymous - **exp_syn**)

Parameters

- **flag_value** (*bool*[451]) – when there's no way to calculate the ratio, the possible cases will be flagged with a negative number. This allows to make substitutions for these values
- **haplotypes** (*bool*[452]) – if true, coverage information is not used, because the SNPs are assumed to come from an alignment that has sequences having haplotypes

Returns

the $\frac{pN}{pS}$ for the gene.

Note: Because pN or pS can be 0, and the return value would be NaN, we take in account some special cases. The default return value in this cases is `numpy.nan`.

- Both synonymous and non-synonymous values are 0:
  - if both the syn and nonsyn attributes are 0 but there's coverage for this gene, we return a 0, as there's no evolution in this gene. Before, the coverage was checked by this method against either the passed *min_cov* parameter that was equal to `MIN_COV`. Now the case is for the user to check the coverage and functions in *mgkit.snps.conv_func* do that. If enough coverage was achieved, the *haplotypes* parameter can be used to return a 0

All other cases return a NaN value

Return type float[453]

calc_ratio_flag()

New in version 0.2.2.

Handles cases where it's important to flag the returned value, as explained in `GeneSNP.calc_ratio()`, and when the both the number of synonymous and non-synonymous is greater than 0, the pN/pS value is returned.

- **The number of non-synonymous is greater than 0 but the number of**

  synonymous is 0:

  - if **flag_value** is **True**, the returned value is **-1**

- The number of synonymous is greater than 0 but the number of non-synonymous is 0:
  - if **flag_value** is **True**, the returned value is **-2**

[451] https://docs.python.org/3/library/functions.html#bool
[452] https://docs.python.org/3/library/functions.html#bool
[453] https://docs.python.org/3/library/functions.html#float

oS	oN	return value
>0	>0	**pN/pS**
0	0	**-3**
>0	0	**-1**
0	>0	**-2**

class mgkit.snps.classes.**SNPType**

Bases: enum.Enum[454]

New in version 0.1.13.

Enum that defines SNP types. Supported at the moment:

- unknown = 0
- syn (synonymous) = 1
- nonsyn (non-synonymous) = 2

Note: No support is planned at the moment to support indel mutations

nonsyn = 2

syn = 1

unknown = 0

mgkit.snps.conv_func module

Wappers to use some of the general function of the snps package in a simpler way.

mgkit.snps.conv_func.**get_full_dataframe**(*snp_data*, *taxonomy*, *min_num=3*, *index_type=None*, *filters=None*)

New in version 0.1.12.

Changed in version 0.2.2: added *filters* argument

Returns a DataFrame with the pN/pS of the given SNPs data.

Shortcut for using *combine_sample_snps()*, using filters from *get_default_filters()*.

Parameters

- **snp_data** (*dict*[455]) – dictionary sample->GeneSyn of SNPs data
- **taxonomy** – Uniprot Taxonomy
- **min_num** (*int*[456]) – minimum number of samples in which a valid pN/pS is found
- **index_type** (*str*[457], *None*) – type of index to return
- **filters** (*iterable*) – list of filters to apply, otherwise uses the default filters

Returns pandas.DataFrame of pN/pS values. The index type is None (gene-taxon)

Return type DataFrame

mgkit.snps.conv_func.**get_gene_map_dataframe**(*snp_data*, *taxonomy*, *gene_map*, *min_num=3*, *index_type='gene'*, *filters=None*)

New in version 0.1.11.

[454] https://docs.python.org/3/library/enum.html#enum.Enum
[455] https://docs.python.org/3/library/stdtypes.html#dict
[456] https://docs.python.org/3/library/functions.html#int
[457] https://docs.python.org/3/library/stdtypes.html#str

Changed in version 0.2.2: added *filters* argument

Returns a `DataFrame` with the pN/pS of the given SNPs data, mapping all taxa to the gene map.

Shortcut for using *combine_sample_snps()*, using filters from *get_default_filters()* and as *gene_func* parameter *map_gene_id()*.

Parameters

- **snp_data** (*dict*[458]) – dictionary sample->GeneSyn of SNPs data
- **taxonomy** – Uniprot Taxonomy
- **min_num** (*int*[459]) – minimum number of samples in which a valid pN/pS is found
- **gene_map** (*dict*[460]) – dictionary of mapping for the gene_ids in in SNPs data
- **index_type** (*str*[461], *None*) – type of index to return
- **filters** (*iterable*) – list of filters to apply, otherwise uses the default filters

Returns `pandas.DataFrame` of pN/pS values. The index type is 'gene'

Return type DataFrame

`mgkit.snps.conv_func.`**`get_gene_taxon_dataframe`**(*snp_data*, *taxonomy*, *gene_map*, *min_num=3*, *rank='genus'*, *index_type=None*, *filters=None*)

New in version 0.1.12.

Changed in version 0.2.2: added *filters* argument

Todo: edit docstring

Returns a `DataFrame` with the pN/pS of the given SNPs data, mapping all taxa to the gene map.

Shortcut for using *combine_sample_snps()*, using filters from *get_default_filters()* and as *gene_func* parameter *map_gene_id()*.

Parameters

- **snp_data** (*dict*[462]) – dictionary sample->GeneSyn of SNPs data
- **taxonomy** – Uniprot Taxonomy
- **min_num** (*int*[463]) – minimum number of samples in which a valid pN/pS is found
- **gene_map** (*dict*[464]) – dictionary of mapping for the gene_ids in in SNPs data
- **index_type** (*str*[465], *None*) – type of index to return
- **filters** (*iterable*) – list of filters to apply, otherwise uses the default filters

Returns `pandas.DataFrame` of pN/pS values. The index type is 'gene'

Return type DataFrame

`mgkit.snps.conv_func.`**`get_rank_dataframe`**(*snp_data*, *taxonomy*, *min_num=3*, *rank='order'*, *index_type='taxon'*, *filters=None*)

New in version 0.1.11.

Changed in version 0.2.2: added *filters* argument

[458] https://docs.python.org/3/library/stdtypes.html#dict
[459] https://docs.python.org/3/library/functions.html#int
[460] https://docs.python.org/3/library/stdtypes.html#dict
[461] https://docs.python.org/3/library/stdtypes.html#str
[462] https://docs.python.org/3/library/stdtypes.html#dict
[463] https://docs.python.org/3/library/functions.html#int
[464] https://docs.python.org/3/library/stdtypes.html#dict
[465] https://docs.python.org/3/library/stdtypes.html#str

Returns a `DataFrame` with the pN/pS of the given SNPs data, mapping all taxa to the specified rank. Higher taxa won't be included.

Shortcut for using *`combine_sample_snps()`*, using filters from *`get_default_filters()`* and as *taxon_func* parameter *`map_taxon_id_to_rank()`*, with include_higher equals to False

Parameters

- **snp_data** (*dict*[466]) – dictionary sample->GeneSyn of SNPs data
- **taxonomy** – Uniprot Taxonomy
- **min_num** (*int*[467]) – minimum number of samples in which a valid pN/pS is found
- **rank** (*str*[468]) – taxon rank to map. Valid ranks are found in `mgkit.taxon.TAXON_RANKS`
- **index_type** (*str*[469], *None*) – type of index to return
- **filters** (*iterable*) – list of filters to apply, otherwise uses the default filters

Returns `pandas.DataFrame` of pN/pS values. The index type is 'taxon'

Return type DataFrame

mgkit.snps.filter module

SNPs filtering functions

`mgkit.snps.filter.`**filter_genesyn_by_coverage**(*gene_syn*, *min_cov=None*)

Checks if the coverage of the provided *gene_syn* is at least *min_cov*

Parameters

- **gene_syn** – `GeneSyn` instance
- **min_cov** (*int*[470]) – minimum coverage allowed (included)

Returns True if the gene has enough coverage

Return type bool[471]

Raises `FilterFails` – if min_cov is None

`mgkit.snps.filter.`**filter_genesyn_by_gene_id**(*gene_syn*, *gene_ids=None*, *exclude=False*, *id_func=None*)

Checks if the gene_id is listed in the filter_list.

Parameters

- **gene_syn** – `GeneSyn` instance
- **gene_ids** (*iterable*) – list of gene IDs to include/exclude
- **exclude** (*bool*[472]) – if the filter is reversed

Returns if the exclude is True, the gene_id must appear in the gene_ids, if False, returns True only if gene_id is NOT in gene_ids.

Return type bool[473]

Raises `FilterFails` – if gene_ids is None

[466] https://docs.python.org/3/library/stdtypes.html#dict
[467] https://docs.python.org/3/library/functions.html#int
[468] https://docs.python.org/3/library/stdtypes.html#str
[469] https://docs.python.org/3/library/stdtypes.html#str
[470] https://docs.python.org/3/library/functions.html#int
[471] https://docs.python.org/3/library/functions.html#bool
[472] https://docs.python.org/3/library/functions.html#bool
[473] https://docs.python.org/3/library/functions.html#bool

mgkit.snps.filter.**filter_genesyn_by_taxon_id**(*gene_syn*, *taxonomy=None*, *filter_list=None*, *exclude=False*, *func=None*)

Checks if the *taxon_id* attribute of *gene_syn* is the *filter_list*. Excelude reverses the result. If func is supplied, it's used to traverse the *taxonomy*.

Parameters

- **gene_syn** – GeneSyn instance
- **taxonomy** – a valid taxonomy (instance of *UniprotTaxonomy*)
- **filter_list** (*iterable*) – list of taxon IDs to include/exclude
- **exclude** (*bool*[474]) – if the filter is reversed
- **func** (*func*) – *is_ancestor()*

Returns if the exclude is True, the gene_id must appear in the gene_ids, if False, returns True only if gene_id is NOT in gene_ids.

Return type bool[475]

Raises FilterFails – if filter_list is None or taxonomy is None and func is not None

mgkit.snps.filter.**get_default_filters**(*taxonomy*, ***kwargs*)

Retuns a list of filters that are used by default. it needs a valid taxonomy and gets the default arguments from mgkit.consts.DEFAULT_SNP_FILTER.

mgkit.snps.filter.**pipe_filters**(*iterable*, **funcs*)

Pipes a list of filter to iterable, using the python ifilter function in the itertools module.

mgkit.snps.funcs module

Functions used in SNPs manipulation

mgkit.snps.funcs.**build_rank_matrix**(*dataframe*, *taxonomy=None*, *taxon_rank=None*)

Make a rank matrix from a pandas.Series with the pN/pS values of a dataset.

Parameters

- **dataframe** – pandas.Series instance with a MultiIndex (gene-taxon)
- **taxonomy** – taxon.UniprotTaxonomy instance with the full taxonomy
- **taxon_rank** (*str*[476]) – taxon rank to limit the specifity of the taxa included

Returns pandas.DataFrame instance

mgkit.snps.funcs.**combine_sample_snps**(*snps_data*, *min_num*, *filters*, *index_type=None*, *gene_func=None*, *taxon_func=None*, *use_uid=False*, *flag_values=False*, *haplotypes=True*)

Changed in version 0.2.2: added *use_uid* argument

Changed in version 0.3.1: added *haplotypes*

Combine a dictionary sample->gene_index->GeneSyn into a pandas.DataFrame. The dictionary is first filtered with the functions in *filters*, mapped to different taxa and genes using *taxon_func* and *gene_func* respectively. The returned DataFrame is also filtered for each row having at least a *min_num* of not NaN values.

Parameters

- **snps_data** (*dict*[477]) – dictionary with the *GeneSNP* instances

[474] https://docs.python.org/3/library/functions.html#bool
[475] https://docs.python.org/3/library/functions.html#bool
[476] https://docs.python.org/3/library/stdtypes.html#str
[477] https://docs.python.org/3/library/stdtypes.html#dict

- **min_num** (*int*[478]) – the minimum number of not NaN values necessary in a row to be returned
- **filters** (*iterable*) – iterable containing filter functions, a list can be found in *mgkit.snps.filter*
- **index_type** (*str*[479], *None*) – if *None*, each row index for the DataFrame will be a MultiIndex with *gene* and *taxon* as elements. If the equals 'gene', the row index will be gene based and if 'taxon' will be taxon based
- **gene_func** (*func*) – a function to map a gene_id to a gene_map. See *mapper.map_gene_id()* for an example
- **taxon_func** (*func*) – a function to map a taxon_id to a list of IDs. See `mapper.map_taxon_id_to_rank` or `mapper.map_taxon_id_to_ancestor` for examples
- **use_uid** (*bool*[480]) – if True, uses the *GeneSNP.uid* instead of *GeneSNP.gene_id*
- **flag_values** (*bool*[481]) – if True, `mgkit.snps.classes.GeneSNP.calc_ratio_flag()` will be used, instead of `mgkit.snps.classes.GeneSNP.calc_ratio()`
- **haplotypes** (*bool*[482]) – if *flag_values* is False, and *haplotypes* is True, the 0/0 case will be returned as 0 instead of NaN

Returns `pandas.DataFrame` with the pN/pS values for the input SNPs, with the columns being the samples.

Return type DataFrame

`mgkit.snps.funcs.`**flat_sample_snps**(*snps_data*, *min_cov*)

New in version 0.1.11.

Adds all the values of a gene across all samples into one instance of `classes.GeneSNP`, giving the average gene among all samples.

Parameters

- **snps_data** (*dict*[483]) – dictionary with the instances of `classes.GeneSNP`
- **min_cov** (*int*[484]) – minimum coverage required for the each instance to be added

Returns the dictionary with only one key (*all_samples*), which can be used with *combine_sample_snps()*

Return type dict[485]

`mgkit.snps.funcs.`**group_rank_matrix**(*dataframe*, *gene_map*)

Group a rank matrix using a mapping, in the form map_id->ko_ids.

Parameters

- **dataframe** – instance of a rank matrix from *build_rank_matrix()*
- **gene_map** (*dict*[486]) – dictionary with the mapping

Returns `pandas.DataFrame` instance

[478] https://docs.python.org/3/library/functions.html#int
[479] https://docs.python.org/3/library/stdtypes.html#str
[480] https://docs.python.org/3/library/functions.html#bool
[481] https://docs.python.org/3/library/functions.html#bool
[482] https://docs.python.org/3/library/functions.html#bool
[483] https://docs.python.org/3/library/stdtypes.html#dict
[484] https://docs.python.org/3/library/functions.html#int
[485] https://docs.python.org/3/library/stdtypes.html#dict
[486] https://docs.python.org/3/library/stdtypes.html#dict

mgkit.snps.funcs.**order_ratios**(*ratios*, *aggr_func=<function median>*, *reverse=False*, *key_filter=None*)

Given a dictionary of id->iterable where iterable contains the values of interest, the function uses aggr_func to sort (ascending by default) it and return a list with the key in the sorted order.

Parameters

- **ratios** (*dict*[487]) – dictionary instance id->iterable
- **aggr_func** (*function*) – any function returning a value that can be used as a key in sorting
- **reverse** (*bool*[488]) – the default is ascending sorting (False), set to True to reverse key_filter: list of keys to use for ordering, if None, every key is used

Returns iterable with the sort order

mgkit.snps.funcs.**significance_test**(*dataframe*, *taxon_id1*, *taxon_id2*, *test_func=<function ks_2samp>*)

New in version 0.1.11.

Perform a statistical test on each gene distribution in two different taxa.

For each gene common to the two taxa, the distribution of values in all samples (columns) between the two specified taxa is tested.

Parameters

- **dataframe** – `pandas.DataFrame` instance
- **taxon_id1** – the first taxon ID
- **taxon_id2** – the second taxon ID
- **test_func** – function used to test, defaults to `scipy.stats.ks_2samp()`

Returns with all pvalues from the tests

Return type pandas.Series

mgkit.snps.funcs.**write_sign_genes_table**(*out_file*, *dataframe*, *sign_genes*, *taxonomy*, *gene_names=None*)

Write a table with the list of significant genes found in a dataframe, the significant gene list is the result of `wilcoxon_pairwise_test_dataframe()`.

Out_file the file name or file object to write the file

Dataframe the dataframe which was tested for significant genes

Sign_genes gene list that are significant

Taxonomy taxonomy object

Gene_names dictionary with the name of the the genes. Optional

mgkit.snps.mapper module

Mapping functions for SNPs - Should be move into an 'iterator' package to be shared with other modules?

mgkit.snps.mapper.**map_gene_id**(*gene_id*, *gene_map=None*)

Returns an iterator for all the values of a dictionary. if gene_id is not found in the gene_map, an empty iterator is returned.

Parameters

- **gene_id** (*immutable*) – gene_id or any other dictionary key.

[487] https://docs.python.org/3/library/stdtypes.html#dict
[488] https://docs.python.org/3/library/functions.html#bool

- **gene_map** (*dict*[489]) – a dictionary in the form key->[v1, v2, .. vN]

Returns iterator (empty if gene_id is not in gene_map) with the values

Return type generator

mgkit.snps.mapper.**map_taxon_id_to_ancestor**(*taxon_id*, *anc_ids=None*, *func=None*)
Given a taxon_id and a list of ancestors IDs, returns an iterator with the IDs that are ancestors of taxon_id.

Parameters

- **taxon_id** (*int*[490]) – taxon ID to be mapped
- **anc_ids** (*iterable*) – taxon IDs to check for ancestry
- **func** – function used to check for ancestry - partial function for *mgkit.taxon.is_ancestor()* that accepts taxon_id and anc_id

Returns iterator with the values or empty

Return type generator

Note: check *mgkit.filter.taxon.filter_taxon_by_id_list()* for examples on using func

---

mgkit.snps.mapper.**map_taxon_id_to_rank**(*taxon_id*, *rank=None*, *taxonomy=None*, *include_higher=False*)
Given a taxon_id, returns an iterator with only the element that correspond to the requested rank. If the taxon returned by *mgkit.taxon.UniprotTaxonomy.get_ranked_taxon* has a different rank than requested, the iterator will be empty if *include_higher* is False and the returned taxon ID if True.

Parameters

- **taxon_id** (*int*[491]) – taxon ID to be mapped
- **rank** (*str*[492]) – taxon rank used (`mgkit.taxon.TAXON_RANKS`)
- **include_higher** (*bool*[493]) – determines if a rank higher than the one requested is to be returned

Returns iterator with the values or empty

Return type generator

Module contents

SNPs data package

mgkit.utils package

Submodules

mgkit.utils.common module

Utility functions

mgkit.utils.common.**apply_func_window**(*func*, *data*, *window*, *step=0*)

mgkit.utils.common.**average_length**(*a1s*, *a1e*, *a2s*, *a2e*)
Given two sets of coordinates, a1 and a2, returns the average length.

[489] https://docs.python.org/3/library/stdtypes.html#dict
[490] https://docs.python.org/3/library/functions.html#int
[491] https://docs.python.org/3/library/functions.html#int
[492] https://docs.python.org/3/library/stdtypes.html#str
[493] https://docs.python.org/3/library/functions.html#bool

Parameters

- **a1s** (*int*[494]) – a1 leftmost number
- **a1e** (*int*[495]) – a1 rightmost number
- **a2s** (*int*[496]) – a2 leftmost number
- **a2e** (*int*[497]) – a2 rightmost number

Return float the average length

mgkit.utils.common.**between**(*pos*, *start*, *end*)
Tests if a number is between two others

Parameters

- **pos** (*int*[498]) – number to test
- **start** (*int*[499]) – leftmost number
- **end** (*int*[500]) – rightmost number

Return bool if the number is between start and end

mgkit.utils.common.**complement_ranges**(*intervals*, *end=None*)
New in version 0.3.1.

Perform a complement operation of the list of intervals, i.e. returning the ranges (tuples) that are not included in the list of intervals. *union_ranges()* is first called on the intervals.

Note: the *end* parameter is there for cases where the ranges passed don't cover the whole space. Assuming a list of ranges from annotations on a nucleotidic sequence, if the last range doesn't include the last position of the sequence, passing end equal to the length of the sequence will make the function include a last range that includes it

---

Parameters

- **intervals** (*intervals*) – iterable where each element is a closed range (tuple)
- **end** (*int*[501]) – if the end of the complement intervals is supposed to be outside the last range.

Returns the list of intervals that complement the ones passed.

Return type list[502]

Examples

```
>>> complement_ranges([(1, 10), (11, 20), (25, 30)], end=100)
[(21, 24), (31, 100)]
>>> complement_ranges([(1, 10), (11, 20), (25, 30)])
[(21, 24)]
>>> complement_ranges([(0, 2), (3, 17), (18, 20)])
[]
```

(continues on next page)

[494] https://docs.python.org/3/library/functions.html#int
[495] https://docs.python.org/3/library/functions.html#int
[496] https://docs.python.org/3/library/functions.html#int
[497] https://docs.python.org/3/library/functions.html#int
[498] https://docs.python.org/3/library/functions.html#int
[499] https://docs.python.org/3/library/functions.html#int
[500] https://docs.python.org/3/library/functions.html#int
[501] https://docs.python.org/3/library/functions.html#int
[502] https://docs.python.org/3/library/stdtypes.html#list

(continued from previous page)

```
>>> complement_ranges([(0, 2), (3, 17), (18, 20)], end=100)
[(21, 100)]
```

mgkit.utils.common.**deprecated**(*func*)
This is a decorator which can be used to mark functions as deprecated. It will result in a warning being emitted when the function is used.

from https://wiki.python.org/moin/PythonDecoratorLibrary

mgkit.utils.common.**range_intersect**(*start1*, *end1*, *start2*, *end2*)
New in version 0.1.13.

Given two ranges in the form *(start, end)*, it returns the range that is the intersection of the two.

Parameters

- **start1** (*int*[503]) – start position for the first range
- **end1** (*int*[504]) – end position for the first range
- **start2** (*int*[505]) – start position for the second range
- **end2** (*int*[506]) – end position for the second range

Returns returns a tuple with the start and end position for the intersection of the two ranges, or *None* if the intersection is empty

Return type (None, tuple[507])

mgkit.utils.common.**range_substract**(*start1*, *end1*, *start2*, *end2*)

mgkit.utils.common.**ranges_length**(*ranges*)
New in version 0.1.12.

Given an iterable where each element is a range, a tuple whose elements are numbers with the first being less than or equal to the second, the function sums the lengths of all ranges.

Parameters **ranges** (*iterable*) – each element is a tuple like *(1, 10)*

Returns sum of all ranges lengths

Return type int[508]

mgkit.utils.common.**union_range**(*start1*, *end1*, *start2*, *end2*)
New in version 0.1.12.

Changed in version 0.3.1: changed behaviour, since the intervals are meant to be closed

If two numeric ranges overlap, it returns the new range, otherwise None is returned. Works on both int and float numbers, even mixed.

Parameters

- **start1** (*numeric*) – start of range 1
- **end1** (*numeric*) – end of range 1
- **start2** (*numeric*) – start of range 2
- **end2** (*numeric*) – end of range 2

Returns union of the ranges or None if the ranges don't overlap

Return type (tuple[509] or None)

[503] https://docs.python.org/3/library/functions.html#int
[504] https://docs.python.org/3/library/functions.html#int
[505] https://docs.python.org/3/library/functions.html#int
[506] https://docs.python.org/3/library/functions.html#int
[507] https://docs.python.org/3/library/stdtypes.html#tuple
[508] https://docs.python.org/3/library/functions.html#int
[509] https://docs.python.org/3/library/stdtypes.html#tuple

Example

```
>>> union_range(10, 13, 1, 10)
(1, 13)
>>> union_range(1, 10, 11, 13)
(1, 13)
>>> union_range(1, 10, 12, 13)
None
```

mgkit.utils.common.**union_ranges**(*intervals*)

New in version 0.3.1.

From a list of ranges, assumed to be closed, performs a union of all elements.

Parameters **intervals** (*intervals*) – iterable where each element is a closed range (tuple)

Returns the list of ranges that are the union of all elements passed

Return type list[510]

Examples

```
>>> union_ranges([(1, 2), (3, 7), (6, 12), (9, 17), (18, 20)])
[(1, 20)]
>>> union_ranges([(1, 2), (3, 7), (6, 12), (9, 14), (18, 20)])
[(1, 14), (18, 20)]
```

mgkit.utils.dictionary module

Dictionary utils

class mgkit.utils.dictionary.**HDFDict**(*file_name*, *table*, *cast=<type 'int'>*)

Bases: object[511]

New in version 0.3.1.

Used a table in a HDFStore (from pandas) as a dictionary. The table must be indexed to perform well. Read only.

Note: the dictionary cannot be modified and exception:*ValueError* will be raised if the table is not in the file

---

mgkit.utils.dictionary.**apply_func_to_values**(*dictionary*, *func*)

New in version 0.1.12.

Assuming a dictionary whose values are iterables, *func* is applied to each element of the iterable, retuning a *set* of all transformed elements.

Parameters

- **dictionary** (*dict*[512]) – dictionary whose values are iterables
- **func** (*func*) – function to apply to the dictionary values

Returns dictionary with transformed values

[510] https://docs.python.org/3/library/stdtypes.html#list
[511] https://docs.python.org/3/library/functions.html#object
[512] https://docs.python.org/3/library/stdtypes.html#dict

Return type dict[513]

class mgkit.utils.dictionary.**cache_dict_file**(*iterator*, *skip_lines=0*)

Bases: object[514]

New in version 0.3.0.

Used to cache the result of a function that yields a tuple (key and value). If the value is found in the internal dictionary (as the class behave), the correspondent value is returned, otherwise the iterator is advanced until the key is found.

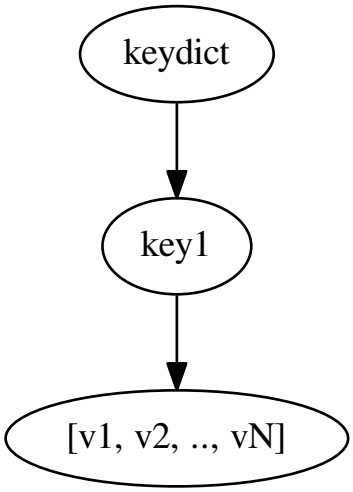
Example

```
>>> from mgkit.io.blast import parse_accession_taxa_table
>>> i = parse_accession_taxa_table('nucl_gb.accession2taxid.gz', key=0)
>>> d = cache_dict_file(i)
>>> d['AH001684']
4400
```

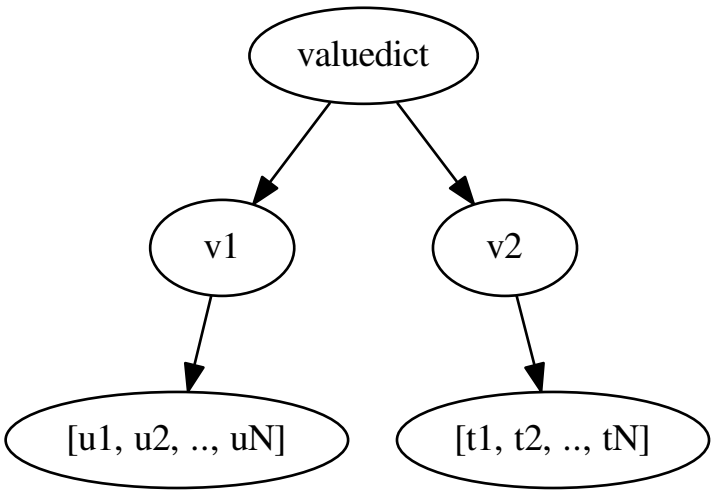
next()

mgkit.utils.dictionary.**combine_dict**(*keydict*, *valuedict*)

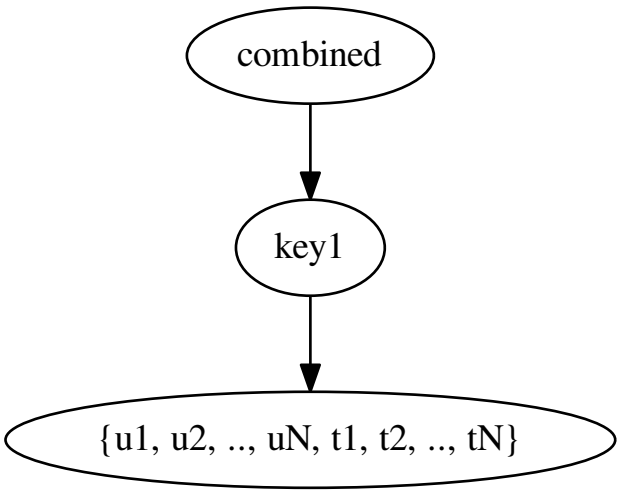
Combine two dictionaries when the values of keydict are iterables. The combined dictionary has the same keys as keydict and the its values are sets containing all the values associated to keydict values in valuedict.



[513] https://docs.python.org/3/library/stdtypes.html#dict
[514] https://docs.python.org/3/library/functions.html#object



Resulting dictionary will be



Parameters

- **keydict** (*dict*[515]) – dictionary whose keys are the same as the returned dictionary
- **valuedict** (*dict*[516]) – dictionary whose values are the same as the returned dictionary

Return dict combined dictionary

mgkit.utils.dictionary.**combine_dict_one_value**(*keydict*, *valuedict*)

Combine two dictionaries by the value of the keydict is used as a key in valuedict and the resulting dictionary

[515] https://docs.python.org/3/library/stdtypes.html#dict
[516] https://docs.python.org/3/library/stdtypes.html#dict

is composed of keydict keys and valuedict values.

Same as `comb_dict()`, but each value in keydict is a single element that is key in valuedict.

Parameters

- **keydict** (*dict*[517]) – dictionary whose keys are the same as the returned dictionary
- **valuedict** (*dict*[518]) – dictionary whose values are the same as the returned dictionary

Return dict combined dictionary

`mgkit.utils.dictionary.`**filter_nan**(*ratios*)
Returns a dictionary with the NaN values taken out

`mgkit.utils.dictionary.`**filter_ratios_by_numbers**(*ratios*, *min_num*)
Returns from a dictionary only the items for which the length of the iterables that is the value of the item, is equal or greater of min_num.

Parameters

- **ratios** (*dict*[519]) – dictionary key->list
- **min_num** (*int*[520]) – minimum number of elements in the value iterable

Return dict filtered dictionary

`mgkit.utils.dictionary.`**find_id_in_dict**(*s_id*, *s_dict*)
Finds a value 's_id' in a dictionary in which the values are iterables. Returns a list of keys that contain the value.

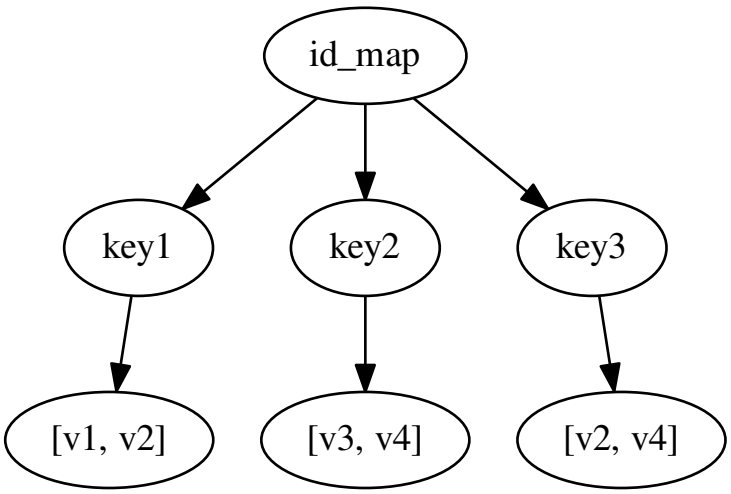
Parameters

- **s_id** (*dict*[521]) – element to look for in the dictionary's values
- **d** (*object*[522]) – dictionary to search in

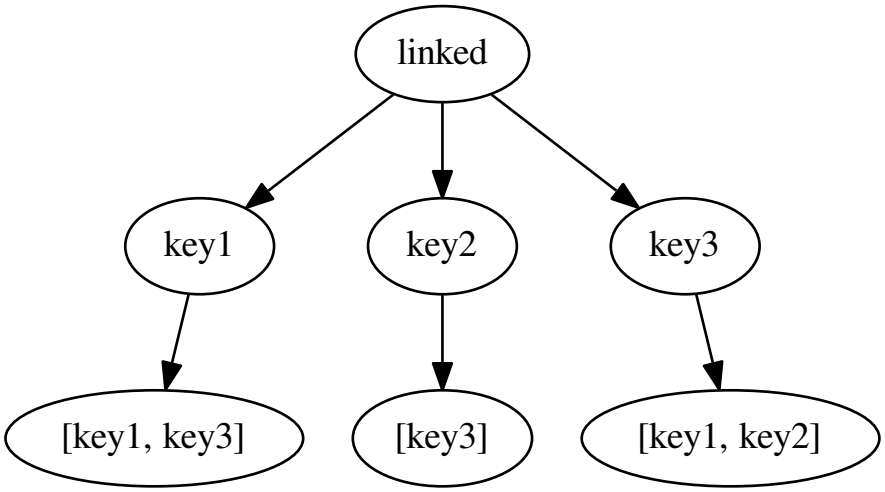
Return list list of keys in which d was found

`mgkit.utils.dictionary.`**link_ids**(*id_map*, *black_list=None*)
Given a dictionary whose values (iterables) can be linked back to other keys, it returns a dictionary in which the keys are the original keys and the values are sets of keys to which they can be linked.

[517] https://docs.python.org/3/library/stdtypes.html#dict
[518] https://docs.python.org/3/library/stdtypes.html#dict
[519] https://docs.python.org/3/library/stdtypes.html#dict
[520] https://docs.python.org/3/library/functions.html#int
[521] https://docs.python.org/3/library/stdtypes.html#dict
[522] https://docs.python.org/3/library/functions.html#object



Becomes:



Parameters

- **id_map** (*dict*[523]) – dictionary of keys to link
- **black_list** (*iterable*) – iterable of values to skip in making the links

Return dict linked dictionary

mgkit.utils.dictionary.**merge_dictionaries**(*dicts*)

New in version 0.3.1.

Merges keys and values from a list/iterable of dictionaries. The resulting dictionary's values are converted into sets, with the assumption that the values are one of the following: float, str, int, bool

[523] https://docs.python.org/3/library/stdtypes.html#dict

mgkit.utils.dictionary.**reverse_mapping**(*map_dict*)

Given a dictionary in the form: key->[v1, v2, .., vN], returns a dictionary in the form: v1->[key1, key2, .., keyN]

Parameters **map_dict** (*dict*[524]) – dictionary to reverse

Return dict reversed dictionary

mgkit.utils.dictionary.**split_dictionary_by_value**(*value_dict*, *threshold*, *aggr_func=<function median>*, *key_filter=None*)

Splits a dictionary, whose values are iterables, based on a threshold:

- one in which the result of aggr_func is lower than the threshold (first)
- one in which the result of aggr_func is equal or greater than the threshold (second)

Parameters

- **valuedict** (*dict*[525]) – dictionary to be splitted
- **threshold** (*number*) – must be comparable to threshold
- **aggr_func** (*func*) – function used to aggregate the dictionary values
- **key_filter** (*iterable*) – if specified, only these key will be in the resulting dictionary

Returns two dictionaries

mgkit.utils.r_func module

mgkit.utils.sequence module

Module containing functions related to sequence data

Note: For those functions without a docstring, look at the same with a underscore ('_') prepended.

class mgkit.utils.sequence.**Alignment**(*seqs=None*)

Bases: object[526]

Simple alignment class

add_seq(*name*, *seq*)

Add a sequence to the alignment

Parameters

- **name** (*str*[527]) – name of the sequence
- **seq** (*str*[528]) – sequence

add_seqs(*seqs*)

Add sequences to the alignment

Parameters **seqs** (*iterable*) – iterable that returns (name, seq)

get_consensus(*nucl=True*)

Changed in version 0.1.16: added *nucl* parameter

[524] https://docs.python.org/3/library/stdtypes.html#dict
[525] https://docs.python.org/3/library/stdtypes.html#dict
[526] https://docs.python.org/3/library/functions.html#object
[527] https://docs.python.org/3/library/stdtypes.html#str
[528] https://docs.python.org/3/library/stdtypes.html#str

The consensus sequence is constructed by checking the nucleotide that has the maximum number of counts for each position in the alignment.

Parameters **nucl** (*bool*[529]) – specify if the alignment is nucleotidic

Returns consensus sequence

Return type str[530]

get_position(*pos*)
Get all characters at a position

Parameters **pos** (*int*[531]) – position to return (0-based)

Return str all characters occuring at the position

get_seq_len()
Get the length of the alignment

get_snps(*ref_seq=None*, *full_size=False*)
A SNP is called for the nucleotide that has the most counts among the ones that differ in the each site of the alignment. If two nucleotides have the same maximum count, one is randomly chosen.

Parameters

- **ref_seq** (*str*[532]) – a reference sequence can be provided, if None, a consensus sequence is produced for the alignment
- **full_size** (*bool*[533]) – if True a tuple is returned for each position in the alignment. If there is no SNP at a position the value for the SNP is None

Return list a list of tuples (position, SNP)

mgkit.utils.sequence.**_get_kmers**(*seq*, *k*)
New in version 0.2.6.

Returns a generator, with every iteration yielding a kmer of size *k*

Parameters

- **seq** (*str*[534]) – sequence
- **k** (*int*[535]) – kmer size

Yields *str* – a portion of *seq*, of size *k* with a step of *1*

mgkit.utils.sequence.**_sequence_signature**(*seq*, *w_size*, *k_size=4*, *step=None*)
New in version 0.2.6.

Returns the signature of a sequence, based on a kmer length, over a sliding window. Each sliding window signature is placed in order into a list, with each element being a collections.Counter[536] instance whose keys are the kmer found in that window.

Parameters

- **seq** (*str*[537]) – sequence for which to get the signature
- **w_size** (*int*[538]) – size of the sliding window size
- **k_size** (*int*[539]) – size of the kmer to use get_kmers()

[529] https://docs.python.org/3/library/functions.html#bool
[530] https://docs.python.org/3/library/stdtypes.html#str
[531] https://docs.python.org/3/library/functions.html#int
[532] https://docs.python.org/3/library/stdtypes.html#str
[533] https://docs.python.org/3/library/functions.html#bool
[534] https://docs.python.org/3/library/stdtypes.html#str
[535] https://docs.python.org/3/library/functions.html#int
[536] https://docs.python.org/3/library/collections.html#collections.Counter
[537] https://docs.python.org/3/library/stdtypes.html#str
[538] https://docs.python.org/3/library/functions.html#int
[539] https://docs.python.org/3/library/functions.html#int

- **step** (*int*[540]) – step to use in `sliding_window()`

Returns a list of `collections.Counter`[541] instances, for each window used

Return type list[542]

`mgkit.utils.sequence._signatures_matrix`(*seqs*, *w_size*, *k_size=4*, *step=None*)
New in version 0.2.6.

Return a matrix (pandas.DataFrame) where the columns are the kmer found in all sequences *seqs* and the rows are the a MultiIndex with the first level being the sequnce name and the second the index of the sliding window for which a signature was computed.

Parameters

- **seqs** (*iterable*) – iterable that yields a tuple, with the first element being the sequence name and the second the sequence itself
- **w_size** (*int*[543]) – size of the sliding window size
- **k_size** (*int*[544]) – size of the kmer to use `get_kmers()`
- **step** (*int*[545]) – step to use in `sliding_window()`, defaults to half of the window size

Returns a DataFrame where the columns are the kmers and the rows are the signatures of each contigs/windows.

Return type pandas.DataFrame

`mgkit.utils.sequence._sliding_window`(*seq*, *size*, *step=None*)
New in version 0.2.6.

Returns a generator, with every iteration yielding a subsequence of size *size*, with a step of *step*.

Parameters

- **seq** (*str*[546]) – sequnece
- **size** (*int*[547]) – size of the sliding window
- **step** (*int*[548], *None*) – the step to use in the sliding window. If *None*, half of the sequence length is used

Yields *str* – a subsequence of size *size* and step *step*

`mgkit.utils.sequence.calc_n50`(*seq_lengths*)
Calculate the N50 statistics for a `numpy.array` of sequence lengths.

The algorithm finds in the supplied array the element (contig length) for which the sum all contig lengths equal or greater than it is equal to half of all assembled base pairs.

Parameters **seq_lengths** (*array*) – an instance of a numpy array containing the sequence lengths

Return int the N50 statistics value

`mgkit.utils.sequence.check_snp_in_seq`(*ref_seq*, *pos*, *change*, *start=0*, *trans_table=None*)
Check a SNP in a reference sequence if it is a synonymous or non-synonymous change.

Parameters

[540] https://docs.python.org/3/library/functions.html#int
[541] https://docs.python.org/3/library/collections.html#collections.Counter
[542] https://docs.python.org/3/library/stdtypes.html#list
[543] https://docs.python.org/3/library/functions.html#int
[544] https://docs.python.org/3/library/functions.html#int
[545] https://docs.python.org/3/library/functions.html#int
[546] https://docs.python.org/3/library/stdtypes.html#str
[547] https://docs.python.org/3/library/functions.html#int
[548] https://docs.python.org/3/library/functions.html#int

- **ref_seq** (*str*[549]) – reference sequence
- **pos** (*int*[550]) – SNP position - it is expected to be a 1 based index
- **change** (*str*[551]) – nucleotide change occuring at *pos*
- **start** (*int*[552]) – the starting position for the coding region - 0 based index
- **trans_table** (*dict*[553]) – translation table used - codon->AA

Return bool True if it is a synonymous change, False if non-synonymous

mgkit.utils.sequence.**convert_aa_to_nuc_coord**(*start*, *end*, *frame=0*)

Converts aa coordinates to nucleotidic ones. The coordinates must be from '+' strand. For the '-' strand, use *reverse_aa_coord()* first.

Parameters

- **start** (*int*[554]) – start of the annotation (lowest number)
- **end** (*int*[555]) – end of the annotation (highest number)
- **frame** (*int*[556]) – frame of the AA translation (0, 1 or 2)

Returns the first element is the converted *start* and the second element is the converted *end*

Return type tuple[557]

Note: the coordinates are assumed to be 1-based indices

---

mgkit.utils.sequence.**get_contigs_info**(*file_name*, *pp=False*)

Changed in version 0.2.4: file_name can be a *dict* name->seq or a list of sequences

New in version 0.2.1.

Given a file name for a fasta file with sequences, a dictionary of name->seq, or a list of sequences, returns the following information in a tuple, or a string if *pp* is True:

- number of sequences
- total base pairs
- max length
- min length
- average length
- N50 statistic

Parameters

- **file_name** (*str*[558]) – fasta file to open
- **pp** (*bool*[559]) – if True, a formatted string is returned

Returns the returned value depends on the value of *pp*, if True a formatted string is returned, otherwise the tuple with all values is.

[549] https://docs.python.org/3/library/stdtypes.html#str
[550] https://docs.python.org/3/library/functions.html#int
[551] https://docs.python.org/3/library/stdtypes.html#str
[552] https://docs.python.org/3/library/functions.html#int
[553] https://docs.python.org/3/library/stdtypes.html#dict
[554] https://docs.python.org/3/library/functions.html#int
[555] https://docs.python.org/3/library/functions.html#int
[556] https://docs.python.org/3/library/functions.html#int
[557] https://docs.python.org/3/library/stdtypes.html#tuple
[558] https://docs.python.org/3/library/stdtypes.html#str
[559] https://docs.python.org/3/library/functions.html#bool

Return type str[560], tuple[561]

mgkit.utils.sequence.**get_seq_expected_syn_count**(*seq*, *start=0*, *syn_matrix=None*)
Calculate the expected number of synonymous and non-synonymous changes in a nucleotide sequence. Assumes that the sequence is already in the correct frame and its length is a multiple of 3.

Parameters

- **seq** (*iterable*) – nucleotide sequence (uppercase chars)
- **start** (*int*[562]) – frame of the sequence
- **syn_matrix** (*dict*[563]) – dictionary that contains the expected number of changes for a codon, as returned by *get_syn_matrix()*

Return tuple tuple with counts of expected counts (syn, nonsyn)

mgkit.utils.sequence.**get_seq_number_of_syn**(*ref_seq*, *snps*, *start=0*, *trans_table=None*)
Given a reference sequence and a list of SNPs, calculates the number of synonymous and non-synonymous SNP.

Parameters

- **ref_seq** (*str*[564]) – reference sequence
- **snps** (*iterable*) – list of tuples (position, SNP) - zero based index
- **start** (*int*[565]) – the frame used for the reference {0, 1, 2}
- **trans_table** (*dict*[566]) – translation table used - codon->AA

Return tuple synonymous and non-synonymous counts

mgkit.utils.sequence.**get_syn_matrix**(*trans_table=None*, *nuc_list=None*)
Returns a dictionary containing the expected count of synonymous and non-synonymous changes that a codon can have if one base is allowed to change at a time.

There are 9 possible changes per codon.

Parameters

- **trans_table** (*dict*[567]) – a tranlation table, defaults to `seq_utils.TRANS_TABLE`
- **nuc_list** (*iterable*) – a list of nucleotides in which a base can change, default to the keys of `seq_utils.REV_COMP`

Return dict returns a dictionary in which for each codon a dictionary {'syn': 0, 'nonsyn': 0} holds the number of expected changes

mgkit.utils.sequence.**get_syn_matrix_all**(*trans_table=None*)
Same as *get_syn_matrix()* but a codon can change in any of the ones included in trans_table.

There are 63 possible changes per codon.

mgkit.utils.sequence.**get_variant_sequence**(*seq*, **snps*)
New in version 0.1.16.

Return a sequence changed in the positions requested.

Parameters

[560] https://docs.python.org/3/library/stdtypes.html#str
[561] https://docs.python.org/3/library/stdtypes.html#tuple
[562] https://docs.python.org/3/library/functions.html#int
[563] https://docs.python.org/3/library/stdtypes.html#dict
[564] https://docs.python.org/3/library/stdtypes.html#str
[565] https://docs.python.org/3/library/functions.html#int
[566] https://docs.python.org/3/library/stdtypes.html#dict
[567] https://docs.python.org/3/library/stdtypes.html#dict

- **seq** (*str*[568]) – a sequence
- ***snps** (*tuple*[569]) – each argument passed is a tuple with the first element as a position in the sequence (1-based index) and the second element is the character to substitute in the sequence

Returns string with the changed characters

Return type str[570]

Example

```
>>> get_variant_sequence('ACTGATATATGCGCGCATCT', (1, 'C'))
'CCTGNTGTATGCGCGCATCT'
```

Note: It is used for nucleotide sequences, but it is valid to use any string

---

mgkit.utils.sequence.**make_reverse_table**(*tbl=None*)

Makes table to reverse complement a sequence by *reverse_complement()*. The table used is the complement for each nucleotide, defaulting to REV_COMP

mgkit.utils.sequence.**put_gaps_in_nuc_seq**(*nuc_seq*, *aa_seq*, *trim=True*)

Match the gaps in an amino-acid aligned sequence to its original nucleotide sequence. If the nucleotide sequence is not a multiple of 3, the trim option by default trim those bases from the output.

Parameters

- **nuc_seq** (*str*[571]) – original nucleotide sequence
- **aa_seq** (*str*[572]) – aligned amino-acid sequence
- **trim** (*bool*[573]) – if True trim last nucleotide(s)

Return str gapped nucleotide sequence

mgkit.utils.sequence.**reverse_aa_coord**(*start*, *end*, *seq_len*)

Used to reverse amino-acid coordinates when parsing an AA annotation on the - strand. Used when the BLAST or HMMER annotations use AA sequences.

Parameters

- **start** (*int*[574]) – start of the annotation
- **end** (*int*[575]) – end of the annotation
- **seq_len** (*int*[576]) – aa sequence length

Returns reversed (from strand - to strand +) coordinates. The first element is the converted *start* and the second element is the converted *end*

Return type tuple[577]

Note:

- start and end are 1-based indices

[568] https://docs.python.org/3/library/stdtypes.html#str
[569] https://docs.python.org/3/library/stdtypes.html#tuple
[570] https://docs.python.org/3/library/stdtypes.html#str
[571] https://docs.python.org/3/library/stdtypes.html#str
[572] https://docs.python.org/3/library/stdtypes.html#str
[573] https://docs.python.org/3/library/functions.html#bool
[574] https://docs.python.org/3/library/functions.html#int
[575] https://docs.python.org/3/library/functions.html#int
[576] https://docs.python.org/3/library/functions.html#int
[577] https://docs.python.org/3/library/stdtypes.html#tuple

mgkit.utils.sequence.**reverse_complement**(*seq*, *tbl='\x00\x01\x02\x03\x04\x05\x06\x07\x08\t\n\x0b\x0c\r\x0e\x0f\x*
!"#$%&\'()+, -*
./0123456789:;<=>?@TBGDEFCHIJKLMNOPQRSAUVWXYZ[\\]^_`)

Returns the reverse complement of a nucleotide sequence

Parameters

- **seq** (*str*[578]) – nucleotide sequence with uppercase characters
- **tbl** (*dict*[579]) – translation table returned by *make_reverse_table()*

Return str returns the reverse complement of a nucleotide sequence

mgkit.utils.sequence.**reverse_complement_old**(*seq*, *tbl=None*)

Returns the reverse complement of a nucleotide sequence

Parameters

- **seq** (*str*[580]) – nucleotide sequence with uppercase characters
- **tbl** (*dict*[581]) – dictionary of complement bases, like REV_COMP

Return str returns the reverse complement of a nucleotide sequence

mgkit.utils.sequence.**sequence_composition**(*sequence*, *chars=('A', 'C', 'T', 'G')*)

New in version 0.1.13.

Returns the number of occurences of each unique character in the sequence

Parameters

- **sequence** (*str*[582]) – sequence
- **chars** (*iterable, None*) – iterable of the chars to test, default to (A, C, T, G). if None checks all unique characters in the sequencce

Yields *tuple* – the first element is the nucleotide and the second is the number of occurences in *sequence*

mgkit.utils.sequence.**sequence_gc_content**(*sequence*)

New in version 0.1.13.

Calculate GC content information for an annotation. The formula is:

$$\frac{(G+C)}{(G+C+A+T)} \tag{7.4}$$

Parameters **sequence** (*str*[583]) – sequence

Returns GC content

Return type float[584]

mgkit.utils.sequence.**sequence_gc_ratio**(*sequence*)

New in version 0.1.13.

Calculate GC ratio information for a sequence. The formula is:

$$\frac{(A+T)}{(G+C)} \tag{7.5}$$

Parameters **sequence** (*str*[585]) – sequence

[578] https://docs.python.org/3/library/stdtypes.html#str
[579] https://docs.python.org/3/library/stdtypes.html#dict
[580] https://docs.python.org/3/library/stdtypes.html#str
[581] https://docs.python.org/3/library/stdtypes.html#dict
[582] https://docs.python.org/3/library/stdtypes.html#str
[583] https://docs.python.org/3/library/stdtypes.html#str
[584] https://docs.python.org/3/library/functions.html#float
[585] https://docs.python.org/3/library/stdtypes.html#str

Returns GC ratio, or *numpy.nan* if G = C = 0

Return type float[586]

mgkit.utils.sequence.**translate_sequence**(*sequence*, *start=0*, *tbl=None*, *reverse=False*)
Translate a nucleotide sequence in an amino acid one.

Parameters

- **sequence** (*str*[587]) – sequence to translate, it's expected to be all caps
- **start** (*int*[588]) – 0-based index for the translation to start
- **tbl** (*dict*[589]) – dictionary with the translation for each codon
- **reverse** (*bool*[590]) – if True, *reverse_complement()* will be called and the returned sequence translated

Return str the translated sequence

mgkit.utils.trans_tables module

The module contains translation tables

Not all genetic codes are included, taken from: http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=t#SG2

Module contents

Package that contains utility functions/classes

mgkit.workflow package

Submodules

mgkit.workflow.add_coverage module

Adds coverage information to GFF

mgkit.workflow.add_coverage.**main**()
Main function

mgkit.workflow.add_coverage.**set_parser**()
argument parser configuration

mgkit.workflow.add_gff_info module

Add more information to GFF annotations: gene mappings, coverage, taxonomy, etc..

[586] https://docs.python.org/3/library/functions.html#float
[587] https://docs.python.org/3/library/stdtypes.html#str
[588] https://docs.python.org/3/library/functions.html#int
[589] https://docs.python.org/3/library/stdtypes.html#dict
[590] https://docs.python.org/3/library/functions.html#bool

Uniprot Command

If the *gene_id* of an annotation is a Uniprot ID, the script queries Uniprot for the requested information. At the moment the information that can be added is the taxon_id, taxon_name, lineage and mapping to EC, KO, eggNOG IDs.

It's also possible to add mappings to other databases using the *-m* option with the correct identifier for the mapping, which can be found at this page[591]; for example if it's we want to add the mappings of uniprot IDs to *BioCyc*, in the *abbreviation* column of the mappings we find that it's identifier is *REACTOME_ID*, so we pass *-m REACTOME* to the script (leaving *_ID* out). Mapped IDs are separated by commas.

The taxonomy IDs are not overwritten if they are found in the annotations, the *-f* is provided to force the overwriting of those values.

See also *MGKit GFF Specifications* for more informations about the GFF specifications used.

Note: As the script needs to query Uniprot a lot, it is recommended to split the GFF in several files, so an error in the connection doesn't waste time.

However, a cache is kept to reduce the number of connections

Taxonomy Command

To refine the taxonomic assignments of predicted genes annotations, the annotation sequences may be searched against a database like the NCBI *nt*.

This commands takes as input a GFF file, one or more blast output files and a file with all mappings from GIDs to taxonomy IDs. More information on how to get the file can be read in the documentation of the function `mgkit.io.blast.parse_gi_taxa_table()`.

The fasta sequences used with BLAST must have as name the uid of the annotations they refer to, and one way to obtain these sequences is to use the function *`mgkit.io.gff.extract_nuc_seqs()`* and save them to a fasta file. Another options is to use the *sequence* command of the *get-gff-info* script (*get-gff-info - Extract informations to GFF annotations*).

The command accept a minimum bitscore to accept an hit and the taxon ID is selected by default using top hit method, but LCA can be used, using the *-l* switch.

Top Hit

The best hit is selected from all those found for a sequence which has the maximum bitscore and identity, with the bitscore having the highest priority.

LCA Taxon

Activated with the *-l* switch, it selects the last common ancestor of all taxon IDs that are from the cellular organism root in the taxonomy and are within a 10 bits (by default, can be customised with *-a*) from the hit with the highest bitscore. If a taxon ID is not found in the taxonomy, it is excluded. One of the requirements of this option is a file that contains the full taxonomy from Uniprot/NCBI. The file can be obtained with the following command:

```
$ download_data -x -p -m your@email
```

The command will output a *taxonomy.pickle* file that can be passed to the *-x* option *download-data - Download Taxonomy from NCBI*.

[591] http://www.uniprot.org/faq/28

Coverage Command

Adds coverage information from BAM alignment files to a GFF file, using the function *mgkit.align.add_coverage_info()*, the user needs to supply for each sample a BAM file, using the *-a* option, whose parameter is in the form *sample,samplealg.bam*. More samples can be supplied adding more *-a* arguments.

Hint: As an example, to add coverage for *sample1*, *sample2* the command line is:

```
add-gff-info coverage -a sample1,sample1.bam -a sample2,sample2.bam \
inputgff outputgff
```

A total coverage for the annotation is also calculated and stored in the *cov* attribute, while each sample coverage is stored into *sample_cov* as per *MGKit GFF Specifications*.

Adding Coverage from samtools depth

The *cov_samtools* allows the use of the output of *samtools* **depth** command. The *-aa* options must be used to pass information about all base pairs and sequences coverage in the BAM/SAM file. The command accept only one sample and the relative file/stream from *samtools*, meaning that multiple samples coverage information must be added one at a time. One solution is to pipe multiple commands to obtain result wanted. For example:

```
$ add-gff-info cov_samtools -s SAMPLE1 -d sample1-coverage input.gff |          add-
↪gff-info cov_samtools -s SAMPLE2 -d sample2-coverage - output.gff
```

This command will add the coverage information for SAMPLE1 and SAMPLE2 from the respective files.

Uniprot Offline Mappings

Similar to the *uniprot* command, it uses the idmapping[592] file provided by Uniprot, which speeds up the process of adding mappings and taxonomy IDs from Uniprot gene IDs. It's not possible tough to add *EC* mappings with this command, as those are not included in the file.

Kegg Information

The *kegg* command allows to add information to each annotation. Right now the information that can be added is restricted to the pathway(s) (reference KO) a KO is part of and both the KO and pathway(s) descriptions. This information is stored in keys starting with **ko_**.

Expected Aminoacidic Changes

Some scripts, like *snp_parser - SNPs analysis*, require information about the expected number of synonymous and non-synonymous changes of an annotation. This can be done using *mgkit.io.gff.Annotation.add_exp_syn_count()* by the user of the command *exp_syn* of this script. The attributes added to each annotation are explained in the *MGKit GFF Specifications*

Adding Information from eggNOG

The *eggnog* command allows to add information from the *annotations* file available for profiles in eggNOG.

[592] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping.dat.gz

Adding Count Data

Count data on a per-sample basis can be added with the *counts* command. The accepted inputs are from HTSeq-count and featureCounts. The ouput produced by featureCounts, is the one from using its **-f** option must be used.

This script accept by default a tab separated file, with a uid in the first column and the other columns are the counts for each sample, in the same order as they are passed to the **-s** option. To use the featureCounts file format, this script **-e** option must be used.

The sample names must be provided in the same order as the columns in the input files. If the counts are FPKMS the *-f* option can be used.

Adding Taxonomy from a Table

There are cases where it may needed or preferred to add the taxonomy from a *gene_id* already provided in the GFF file. For such cases the *addtaxa* command can be used. It works in a similar way to the *taxonomy* command, only it expect three different type of inputs:

- *GI-Taxa* table from NCBI (e.g. gi_taxid_nucl.dmp,)
- tab separated table
- dictionary
- HDF5

The first two are tab separated files, where on each line, the first column is the *gene_id* that is found in the first column, while the second if the *taxon_id*.

The third option is a serialised Python *dict*/hash table, whose keys are the *gene_id* and the value is that gene corresponding *taxon_id*. The serialised formats accepted are msgpack, json and pickle. The *msgpack* module must be importable. The option to use json and msgpack allow to integrate this script with other languages without resorting to a text file.

The last option is a HDF5 created using the *to_hdf* command in *taxon-utils - Taxonomy Utilities*. This requires *pandas* installed and *pytables* and it provides faster lookup of IDs in the table.

While the default is to look for the *gene_id* attribute in the GFF annotation, another attribute can be specified, using the **-gene-attr** option.

Note: the dictionary content is loaded after the table files and its keys and corresponding values takes precedence over the text files.

Warning: from September 2016 NCBI will retire the GI. In that case the same kind of table can be built from the *nucl_gb.accession2taxid.gz* file The format is different, but some information can be found in `mgkit.io.blast.parse_accession_taxa_table()`

Adding information from Pfam

Adds the Pfam description for the annotation, by downloading the list from Pfam.

The options allow to specify in which attribute the ID/ACCESSION is stored (defaults to *gene_id*) and which one between ID/ACCESSION is the value of that attribute (defaults to *ID*). if no description is found for the family, a warning message is logged.

Changes

Changed in version 0.3.0: added *cov_samtools* command, *–split* option to *exp_syn*, *-c* option to *addtaxa*

Changed in version 0.2.6: added *skip-no-taxon* option to *addtaxa*

Changed in version 0.2.5: if a dictionary is supplied to *addtaxa*, the GFF is not preloaded

Changed in version 0.2.3: added *pfam* command, renamed *gitaxa* to *addtaxa* and made it general

Changed in version 0.2.2: added *eggnog*, *gitaxa* and *counts* command

Changed in version 0.2.1.

- added *-d* to *uniprot* command
- added cache to *uniprot* command
- added *kegg* command (cached)

Changed in version 0.1.16: added *exp_syn* command

Changed in version 0.1.15: *taxonomy* command *-b* option changed

Changed in version 0.1.13.

- added *–force-taxon-id* option to the *uniprot* command
- added *coverage* command
- added *taxonomy* command
- added *unipfile* command

New in version 0.1.12.

mgkit.workflow.add_gff_info.**add_uniprot_info**(*annotations*, *options*, *info_cache*)

mgkit.workflow.add_gff_info.**addtaxa_command**(*options*)

mgkit.workflow.add_gff_info.**choose_by_lca**(*hits*, *taxonomy*, *gid_taxon_map*, *score=10*)

mgkit.workflow.add_gff_info.**choose_by_score**(*hits*, *gid_taxon_map*)

mgkit.workflow.add_gff_info.**counts_command**(*options*)

mgkit.workflow.add_gff_info.**coverage_command**(*options*)

mgkit.workflow.add_gff_info.**eggnog_command**(*options*)

mgkit.workflow.add_gff_info.**exp_syn_command**(*options*)

mgkit.workflow.add_gff_info.**get_gids**(*uid_gid_map*)

mgkit.workflow.add_gff_info.**kegg_command**(*options*)

mgkit.workflow.add_gff_info.**load_counts**(*count_files*, *samples*, *featureCounts*)

mgkit.workflow.add_gff_info.**main**()
Main function

mgkit.workflow.add_gff_info.**parse_hdf5_arg**(*argument*)

mgkit.workflow.add_gff_info.**pfam_command**(*options*)

mgkit.workflow.add_gff_info.**read_lines_from_files**(*file_handles*)

mgkit.workflow.add_gff_info.**samtools_depth_command**(*options*)

mgkit.workflow.add_gff_info.**set_addtaxa_parser**(*parser*)

mgkit.workflow.add_gff_info.**set_blast_taxonomy_parser**(*parser*)

mgkit.workflow.add_gff_info.**set_common_options**(*parser*)

mgkit.workflow.add_gff_info.**set_counts_parser**(*parser*)

mgkit.workflow.add_gff_info.**set_coverage_parser**(*parser*)

mgkit.workflow.add_gff_info.**set_eggnog_parser**(*parser*)

mgkit.workflow.add_gff_info.**set_exp_syn_parser**(*parser*)
New in version 0.1.16.

mgkit.workflow.add_gff_info.**set_kegg_parser**(*parser*)

mgkit.workflow.add_gff_info.**set_parser**()
Sets command line arguments parser

mgkit.workflow.add_gff_info.**set_pfam_parser**(*parser*)

mgkit.workflow.add_gff_info.**set_samtools_depth_parser**(*parser*)

mgkit.workflow.add_gff_info.**set_uniprot_offline_parser**(*parser*)

mgkit.workflow.add_gff_info.**set_uniprot_parser**(*parser*)

mgkit.workflow.add_gff_info.**split_sample_alg**(*argument*)
Split sample/alignment option

mgkit.workflow.add_gff_info.**taxonomy_command**(*options*)

mgkit.workflow.add_gff_info.**uniprot_command**(*options*)

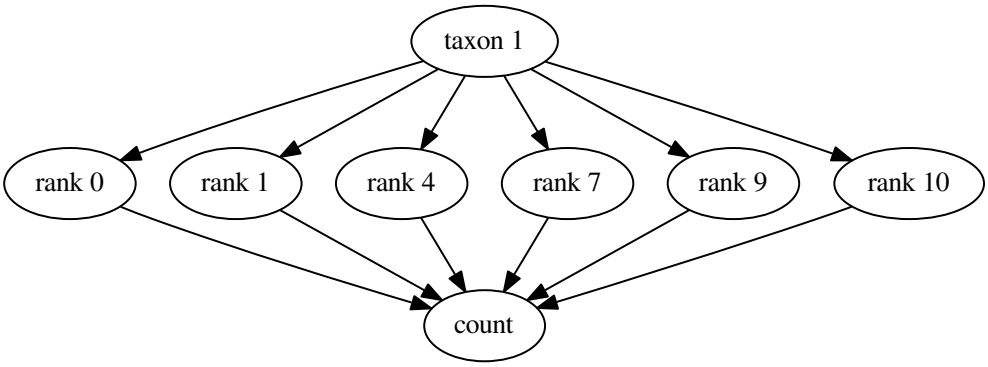
mgkit.workflow.add_gff_info.**uniprot_offline_command**(*options*)

mgkit.workflow.assembly module

Workflow associated with assembly statistics and evaluation

mgkit.workflow.assembly.**assign_contigs_to_taxa**(*annotations*, *root_map=None*, *black_list=None*)
Groups annotations by contig (seq_id) and counts how many contigs a taxon, or its root if root_map is supplied, have been assigned to.

The actual form of the dictionary like this:



Note: the number of ranks for a taxon is not pretedermined, but depends on the values returned by *rank_annotations_by_attr()*.

Parameters

- **annotations** (*iterable*) – list of gff.GFFKegg instances

- **root_map** (*dict*[593]) – dictionary taxon->root

Return dict dictionary

mgkit.workflow.assembly.**basic_stats**(*array*, *sep*)
Returns formatted basic statistics for contig lengths

mgkit.workflow.assembly.**filter_contig_assignments**(*contig_assign*, *threshold=5*, *min_counts=1*)
Filter contigs assignments using a threshold for the rank: all rank counts belonging to a taxon which are greater than or equal to threshold will be summed up.

Parameters

- **contig_assign** (*dict*[594]) – dictionary returned by *assign_contigs_to_taxa()*
- **threshold** (*int*[595]) – the minimum rank for which the counts are summed up

Return dict dictionary in the form taxon_name->count

mgkit.workflow.assembly.**rank_annotations_by_attr**(*annotations*, *attr='taxon'*)
For all annotations in the list (usually all annotations for a contig), counts how many time a set attribute 'attr' appears. The resulting dictionary is then sorted by the number of counts and the one with the highest count is ranked by how much it represent the total number of counts.

The rank is an integer number between 0 and 10.

Parameters

- **annotations** (*iterable*) – list of gff.GFFKegg instances
- **attr** (*str*[596]) – the attribute for which the annotations are counted

Return tuple the attr with the most counts and its rank

mgkit.workflow.assembly.**write_fasta_summary**(*file_handle*, *seq_lengths*, *seq_lengths_filt*, *sep='\t'*)
Write summary file for assembly

Parameters

- **file_handle** – file handle for output
- **seq_lengths** (*array*) – array for sequence lengths
- **seq_lengths_filt** (*array*) – array for sequence lengths of annotated contigs
- **sep** (*str*[597]) – string used as column separator

mgkit.workflow.blast2gff module

Blast output conversion in GFF requires a BLAST+ tabular format which can be obtained by using the *–outfmt 6* option with the default columns, as specified in *mgkit.io.blast.parse_blast_tab()*. The script can get data from the standard in and ouputs GFF lines on the standard output by default.

Uniprot

The Function *mgkit.io.blast.parse_uniprot_blast()* is used, which filters BLAST hits based on bitscore and adds by default a *db* attribute to the annotation with the value *UNIPROT-SP*, indicating that the SwissProt db is used and a *dbq* attribute with the value 10. The feature type used in the GFF is CDS.

[593] https://docs.python.org/3/library/stdtypes.html#dict
[594] https://docs.python.org/3/library/stdtypes.html#dict
[595] https://docs.python.org/3/library/functions.html#int
[596] https://docs.python.org/3/library/stdtypes.html#str
[597] https://docs.python.org/3/library/stdtypes.html#str

BlastDB

If a BlastDB, such as *nt* or *nr* was used, the **blastdb** command offers some quick defaults to parse BLAST results.

It now includes options to control the way the sequence header are formatted. Options to change the separator used, as well as the column used as *gene_id*. This was added because at the moment the GI identifier (the second column in the header) is used, but it's being phased out in favour of the embl/gb/dbj (right now the fourth column in the header). This should easy the transition to the new format and makes it easier to adapt an older pipeline/blastdb to newer files (like the ID to TAXA files).

The header from the a *ncbi-nt* header looks like this:

```
gi|160361034|gb|CP000884.1
```

This is the default output accepted by the *blastdb* command. The fields are separated by | (pipe) and the GI is used (–gene-index 1, since internally the string is split by the separator and the second element is take - lists indices are 0-based in Python). This output uses the following options:

```
--header-sep '|' --gene-index 1
```

Notice the single quotes to pass the pipe symbol, since *bash* would interpret it as pipeing to the next coommand otherwise. This is the default.

In case, for the same header, we want to use the *gb* identifier, the only option to be specified is:

```
--gene-index 3
```

This will get the fourth element of the header (since we're splitting by pipe).

As in the *uniprot* command, the *gene_id* can be set to use the whole header, using the *-n* option. Useful in case the *BLAST* db that was used was custom made. While pipe is used in major databases, it was made the default, by if the db used has different conventions the separator can be changed. There's also the options of later changing the *gene_id* in the output GFF if necessary.

Changes

Changed in version 0.2.6: added *-r* option to *blastdb*

Changed in version 0.2.5: added more options to give user control to the *blastdb* command

New in version 0.2.3: added *–fasta-file* option, added more data from a blsat hit

New in version 0.2.2: added *blastdb* command

Changed in version 0.2.1: added *-ft* option

Changed in version 0.1.13.

- added *-n* parameter to *uniprot* command
- added *-k* option to *uniprot* command

New in version 0.1.12.

`mgkit.workflow.blast2gff.`**`convert_from_blastdb`**`(`*options*`)`
New in version 0.2.2.

mgkit.workflow.blast2gff.**convert_from_uniprot**(*options*)

mgkit.workflow.blast2gff.**load_fasta_file**(*file_name*)

mgkit.workflow.blast2gff.**main**()
Main function

mgkit.workflow.blast2gff.**parse_attr_arg**(*value*)

mgkit.workflow.blast2gff.**set_blastdb_parser**(*parser*)
New in version 0.2.2.

mgkit.workflow.blast2gff.**set_common_options**(*parser*)

mgkit.workflow.blast2gff.**set_parser**()
Changed in version 0.2.2: added *blastdb* command

Sets command line arguments parser

mgkit.workflow.blast2gff.**set_uniprot_parser**(*parser*)

mgkit.workflow.download_data module

The scripts downloads the data that is used by the framework for some of its functions. It's mostly a shortcut to call the download_data function that is present in every module that is in the package mappings and in the kegg module.

The only option required, is the email contact for the person using the script; this is used to make sure that the API requirements in Uniprot are fullfilled and they can contact the person using the script is any problem arise.

Note: The default behavior is to download first the taxonomy data form Uniprot, Kegg and additional mapping data.

Taxonomy

It is downloaded from Uniprot and build a data structure that is used by several scripts and function in the package. The download can take some time.

Warning: if only the taxonomy is to be downloaded, both the *-x* and *-p* options must be passed to the script.

Kegg Onthologs

The Kegg data is the only "required" data at the moment, because it's used to download the sequence data (via the donwload_profiles script) for the profiling. It is the only data that can't be saved unless it's fully downloaded.

Kegg data is required by the mappers currently supported, and its download takes longer. The mappers handle timeouts and if exceptions are raised the data is saved and the download is resumed when the script is started again.

Other Mappings

The other mappings (from KO) are downloaded by default and this can be excluded by using the *-p* option. Mappings for Gene Onthology, eggNOG and CaZy are downloaded.

As the download of the mappings can take a lot of time, or break because of the number of requests to the web sites, checkpoints are saved often, so it can resumed at a later time.

Warning: the Gene Onthology module has specific requirements, so if they are not downloaded

mgkit.workflow.download_data.**main**()
Main function

mgkit.workflow.download_data.**set_parser**()
Sets command line arguments parser

mgkit.workflow.download_profiles module

Overview

This script downloads sequence data for each gene of interest (ortholog) and all the specified taxa. The files that are downloaded with this script can then be used to create HMMER profiles, to search for similarity in a aminoacidic or nucleotidic sequence.

Limitations

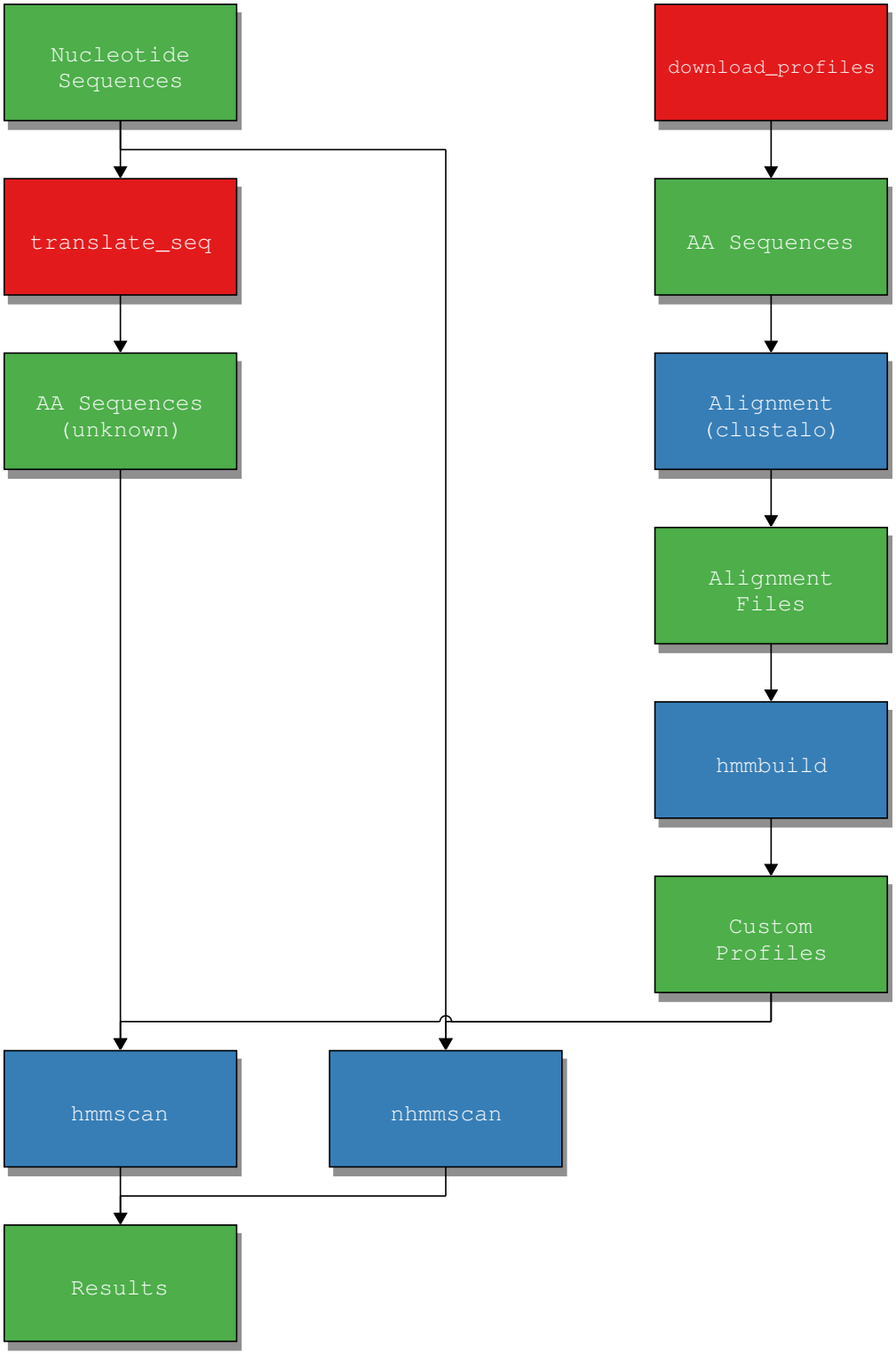
At the moment, the script uses Kegg Orthologs as the ortholog database.

Warning: Some taxa may still black listed, because they are not relevant to the rumen microbiome. If you find such thing to occur to you, please contact me or open an issue on the repository.

Required Data

The script requires data from Kegg Orthologs and Uniprot to be downloaded, before it can be used. The script **download_data** (*download-data - Download Taxonomy from NCBI*) automates the process.

Workflow for Custom Profiles



The process of building the profiles to be used with HMMER is a step that involves several tasks (illustrated in the Workflow above):

1. download of data

2. alignment of sequences
3. conversion in HMMER profiles.

The first step involves, for all ortholog genes, to download all sequences available for each taxon level of interest: this will produce a series of file which contain the amino-acid sequences for each tuple gene-taxon. This sctipt, *download_profiles* can be used. The aminoacidic sequences downloaded are then aligned using Clustal Omega (or other) and for each alignment a profile is built.

HMMER required the use of aminoacidic sequences, to be match against the profiles. The *translate_seq* script can be used to translate nucleotidic sequences into aminoacidic ones. However, the last version of HMMER should be able to match nucleotidic sequences, but it was not tested by us. The example Workflow above illustrate that.

Building profiles in this way, by going through all ortholog genes and choosing the taxon level desired, opens the possibility of incrementally refining the profiling of a metagenome without having to rerun all profiles again, as only the new ones need to be run. Filtering the all the results is a much faster operation.

Usage

The default behaviour is to download all Kegg Orthologs for all taxa in the given taxonomy. Taxa can be filtered by both lineage (e.g. archaea, carnivora, etc.) and rank (e.g. genus, family, etc.). Another option is to specify the KO and taxa IDs to download.

Taxa Filters

The way a taxon is specified is through a few different rules:

- specific **taxon ids** in uniprot
- a specific **taxon rank** (e.g.: genus, phylum, etc.)
- optional lineage filter: the lineage filter make sure that the name specified is included in the lineage attribute in the taxonomy.

As an example, if the rank chosen is genus, and the lineage option is set to archaea, only the taxa whose rank is genus and that belong to the archaea subtree will be downloaded:

```
$ download_profiles -m EMAIL -r genus -l archaea mg_data/kegg.pickle \
-t mg_data/taxonomy.pickle
```

This allows to customise the level of specificity that we want in profiling and make the process of downloading faster. For metagenomic data, a good start is mixing different taxon ranks, using the order or genus for the genes and then specifying a lineage of interest.

Because each profile is indipendent from each other, it's useful to start the download with a certain rank and then run the profiling. During the profiling a new download can be started and so on.

Specific Genes and Taxa

It is possible to download only specific taxa and KO and can be done using the *-i* and *-ko* respectively. When *-ko* is used, loading Kegg Data with *-k* is not required and it is up to the user to ensure the correct genes or taxa.

An example to download only KO from 3 different taxa:

```
$ download_profiles -v -m EMAIL -ko K00016 -i 9611 9645 9682 \
-t mg_data/taxonomy.pickle
```

The same example using taxa filtering, instead (at the time of writing):

```
$ download_profiles -v -m EMAIL -ko K00016 -r genus -l carnivora \
-t mg_data/taxonomy.pickle
```

Changes

Changed in version 0.2.1: added *-ko* option, resolved issues caused by changes in library

mgkit.workflow.download_profiles.**add_profiles_to_length**(*seqs*, *length_data*)
Adds the average profile length to the dictionary

mgkit.workflow.download_profiles.**choose_ko_ids**(*kegg_data*, *options*)
Returns the list of mapping ID->Name according to the options passed

mgkit.workflow.download_profiles.**choose_taxa**(*taxonomy*, *options*)
Returns the list of ids to look for in Uniprot

mgkit.workflow.download_profiles.**download_ko_sequences**(*ko_id*, *taxon_ids*, *reviewed*, *contact*)
Downloads the sequences associated to all taxon IDs provided

mgkit.workflow.download_profiles.**filter_found_taxa**(*taxon_ids_found*, *taxon_ids*, *taxonomy*)
Filter the taxa found in Uniprot, making sure that they at a lower level of those requested

mgkit.workflow.download_profiles.**filter_taxonomy_by_lineage**(*taxonomy*, *taxon_ids*, *lineage*)

mgkit.workflow.download_profiles.**filter_taxonomy_by_rank**(*taxonomy*, *taxon_ids*, *rank*)

mgkit.workflow.download_profiles.**load_data**(*taxon_data*, *length_data_name*)
Loads data for script

mgkit.workflow.download_profiles.**main**()
Main function

mgkit.workflow.download_profiles.**map_ko_to_uniprot**(*ko_id*, *taxon_ids*, *reviewed*, *contact*)
Returns the taxon IDs found in Uniprot for a specific id

mgkit.workflow.download_profiles.**set_parser**()
argument parser configuration

mgkit.workflow.download_profiles.**write_ko_sequences**(*seqs*, *taxonomy*, *output_dir*)
Writes fasta sequences to disc

mgkit.workflow.extract_gff_info module

mgkit.workflow.fasta_utils module

New in version 0.3.0.

Scripts that includes some functionality to help use FASTA files with the framework

split command

Used to split a fasta file into smaller fragments

translate command

Used to translate nucleotide sequences into amino acids.

uid command

Used to change a FASTA file headers to a unique ID. A table (tab separated) with the changes made can be kept, using the *–table* option.

Changes

New in version 0.3.0.

Changed in version 0.3.1: added *translate* and *uid* command

mgkit.workflow.fasta_utils.**load_trans_table**(*table_name*)
Loads translation table

mgkit.workflow.fasta_utils.**main**()
Main function

mgkit.workflow.fasta_utils.**set_parser**()
Sets command line arguments parser

mgkit.workflow.fasta_utils.**set_split_parser**(*parser*)

mgkit.workflow.fasta_utils.**set_translate_parser**(*parser*)

mgkit.workflow.fasta_utils.**set_uid_parser**(*parser*)

mgkit.workflow.fasta_utils.**split_command**(*options*)

mgkit.workflow.fasta_utils.**translate_command**(*options*)

mgkit.workflow.fasta_utils.**translate_seq**(*name*, *seq*, *trans_table*)
Tranlates sequence into the 6 frames

mgkit.workflow.fasta_utils.**uid_command**(*options*)

mgkit.workflow.fastq_utils module

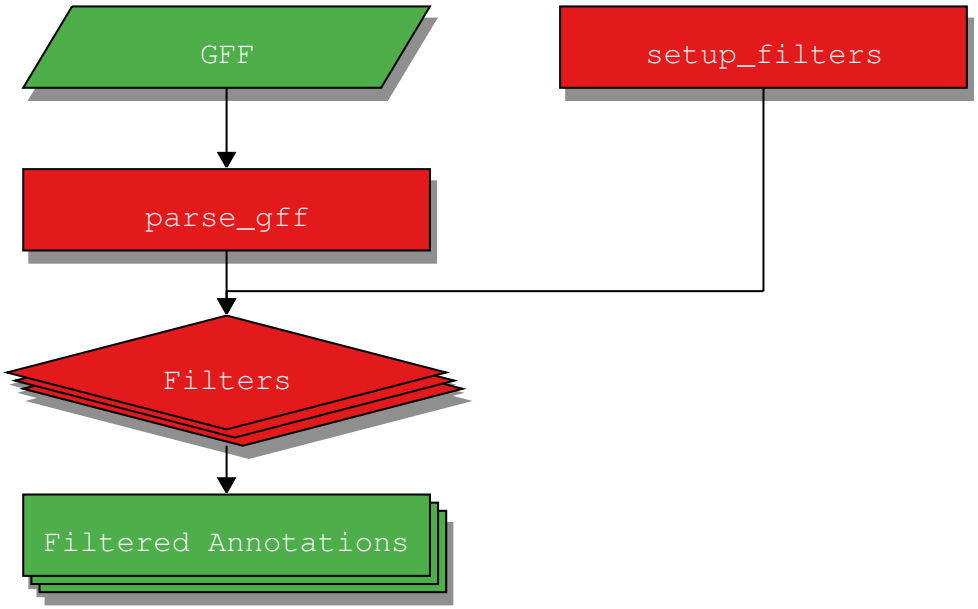
mgkit.workflow.filter_gff module

Filters GFF annotations in different ways.

Value Filtering

Enables filtering of GFF annotations based on the the first 8 columns, which are fixed values as well using the last column which holds information in a key=value way. There are some predefined key=value filters, like *gene_id*, but *–str-eq*, *–str-in*, *–num-ge* and *–num-le* allow to make additional filters.

The functions used to make the filters are located in the module *mgkit.filter.gff*, and their names start with *filter_base*, *filter_attr* and *filter_len*.



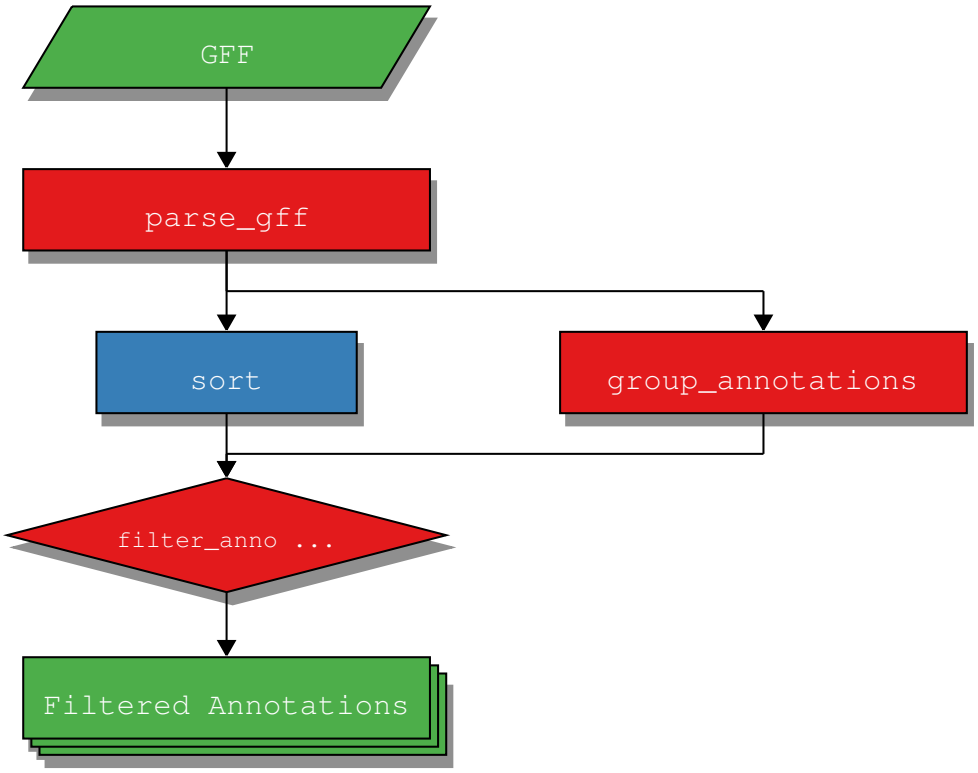
Overlap Filtering

Filters overlapping annotations using the functions *mgkit.filter.gff.choose_annotation()* and *mgkit.filter.gff.filter_annotations()*, after the annotations are grouped by both sequence and strand. If the GFF is sorted by sequence name and strand, the *-t* can be used to make the filtering use less memory. It can be sorted in Unix using *sort -s -k 1,1 -k 7,7 gff_file*, which applies a stable sort using the sequence name as the first key and the strand as the second key.

Note: It is also recommended to use:

```
export LC_ALL=C
```

To speed up the sorting



The above digram describes the internals of the script.

The annotations needs first to be grouped by seq_id and strand, forming a group that can be then be passed to `mgkit.filter.gff.filter_annotations()`. This function:

1. sort annotations by bit score, from the highest to the lowest
2. loop over all combination of N=2 annotations:
   (a) choose which of the two annotations to discard if they overlap for a the required amount of bp (defaults to 100bp)
   (b) in which case, the preference is given to the db quality first, than the bit score and finally the lenght of annotation, the one with the highest values is kept

While the default behaviour is the same, now it is posible to decided the function used to discard one the two annotations. It is possible to use the *-c* argument to pass a string that defines the function. The string passed must start with or without a **+**. Using **+** translates into the builtin function *max* while no **+** translates into *min* from the second character on, any number of attributes can be used, separated by commas. The attributes, however, must be one of the properties defined in `mgkit.io.gff.Annotation`, *bitscore* that returns the value converted in a *float*. Internally the attributes are stored as strings, so for attributes that have no properties in the class, such as *evalue*, the *float* builtin is applied.

The tuples built for both annotations are then passed to the comparison function to be selected and the value returned by it is **discarded**. The order of the elements in the string is important to define the priority given to each element in the comparison and the leftmost one has the highesst priority.

Examples of function strings:

- *-dbq,bitscore,length* becomes max((ann1.dbq, ann1.bitscore, ann1.length), (ann2.dbq, ann2.bitscore, ann2.length) - This is default and previously only choice
- *-bitscore,length,dbq* uses the same elements but gives lowest priority to *dbq*
- *+evalue*: will discard the annotation with the highest *evalue*

Per Sequence Values

The *sequence* command allows to filter on a per sequence basis, using functions such as the median, quantile and mean on attributes like evalue, bitscore and identity. The file can be passed as sorted already, saving memory (like in the *overlap* command), but it's not needed to sort the file by strand, only by the first column.

Coverage Filtering

The *cov* command calculates the coverage of annotations as a measure of the percentage of each reference sequence length. A minimum coverage percentage can be used to keep the annotations of sequences that have a greater or equal coverage than the specified one.

Changes

New in version 0.1.12.

Changed in version 0.1.13: added *–sorted* option

Changed in version 0.2.0: changed option *-c* to accept a string to filter overlap

Changed in version 0.2.5: added *sequence* command

Changed in version 0.2.6: added *length* as attribute and *min*/*max*, and *ge* is the default comparison for command *sequence*, *–sort-attr* to *overlap*

Changed in version 0.3.1: added *–num-gt* and *–num-lt* to *values* command, added *cov* command

mgkit.workflow.filter_gff.**common_options**(*parser*)

mgkit.workflow.filter_gff.**coverage_command**(*options*)

mgkit.workflow.filter_gff.**filter_overlaps**(*options*)

mgkit.workflow.filter_gff.**filter_perseq**(*options*)

mgkit.workflow.filter_gff.**filter_values**(*options*)

mgkit.workflow.filter_gff.**find_comparison**(*options*)

mgkit.workflow.filter_gff.**find_function**(*options*)

mgkit.workflow.filter_gff.**main**()
Main function

mgkit.workflow.filter_gff.**make_choose_func**(*argument*)
Builds the function used to choose between two annotations.

mgkit.workflow.filter_gff.**parse_attr_arg**(*value*, *convert=<type 'str'>*)

mgkit.workflow.filter_gff.**perseq_calc_threshold**(*annotations*, *attribute*, *function*, *func_arg=None*)

mgkit.workflow.filter_gff.**set_coverage_parser**(*parser*)

mgkit.workflow.filter_gff.**set_overlap_parser**(*parser*)

mgkit.workflow.filter_gff.**set_parser**()
Sets command line arguments parser

mgkit.workflow.filter_gff.**set_perseq_parser**(*parser*)

mgkit.workflow.filter_gff.**set_values_parser**(*parser*)

mgkit.workflow.filter_gff.**setup_filters**(*options*)

mgkit.workflow.hmmer2gff module

Script to convert HMMER results files (domain table) to a GFF file, the name of the profiles are expected to be now in the form *GENEID_TAXONID_TAXON-NAME(-nr)* by default, but any other profile name is accepted.

The profiles tested are those made from Kegg Orthologs, from the *download_profiles* script. If the *–no-custom-profiles* options is used, the script can be used with any profile name. The profile name will be used for *gene_id*, *taxon_id* and *taxon_name* in the GFF file.

It is possible to use seuqnces not translated using mgkit, no information on the frame is assumed, so this script can be used against a protein DB. For example Uniprot can be searched for profiles, in which case the **–no-frame** options must be used.

Note: for GENEID, old documentation points to KOID, it is the same

Warning: The compatibility with old data has been **removed**, meaning that old experiments must use the scripts from those versions. It is possible to use multiple environments, with *virtualenv* for this purpose. An examples is given in *Installation*.

Changes

Changed in version 0.1.15: adapted to new GFF module and specs

Changed in version 0.2.1: added options to customise output and filters and old restrictions

Changed in version 0.3.1: added *–no-frame* option for non mgkit-translated proteins, sequence headers are handled the same way as HMMER (truncated at the first space)

`mgkit.workflow.hmmer2gff.`**`get_aa_data`**(*f_handle*)
Load aminoacid seuqnces used by HMMER.

`mgkit.workflow.hmmer2gff.`**`main`**()
Main loop

`mgkit.workflow.hmmer2gff.`**`parse_domain_table_contigs`**(*options*)
Parse the HMMER result file

`mgkit.workflow.hmmer2gff.`**`set_parser`**()
Setup command line options

mgkit.workflow.json2gff module

New in version 0.2.6.

This script converts annotations in JSON format that were created using MGKit back into GFF annotations.

mongodb command

Annotations converted into MongoDB records with *get-gff-info mongodb* can be converted back into a GFF file using this command. It can be useful to get a GFF file as output from a query to a MongoDB instance on the command line.

For example:

mongoexport -d db -c test | json2gff mongodb

will convert all the annotations in the database *db*, collection *test* to the standard out.

mgkit.workflow.json2gff.**main**()
Main function

mgkit.workflow.json2gff.**mongodb_command**(*options*)

mgkit.workflow.json2gff.**set_common_options**(*parser*)

mgkit.workflow.json2gff.**set_mongodb_parser**(*parser*)

mgkit.workflow.json2gff.**set_parser**()
Sets command line arguments parser

mgkit.workflow.nuc2aa module

Translate nucleotidic sequence in amino acidic sequences

mgkit.workflow.nuc2aa.**load_trans_table**(*table_name*)
Loads translation table

mgkit.workflow.nuc2aa.**main**()
Main function

mgkit.workflow.nuc2aa.**set_parser**()
argument parser configuration

mgkit.workflow.nuc2aa.**translate_buff**(*jobs*, *buff_seqs*, *trans_table*)
Process a set amount of sequences - multiprocess

mgkit.workflow.nuc2aa.**translate_seq**(*name*, *seq*, *trans_table*)
Tranlate sequence into the 6 frames

mgkit.workflow.sampling_utils module

New in version 0.3.1.

Resampling Utilities

sample command

This command samples from a Fasta or FastQ file, based on a probability defined by the user (0.001 or 1 / 1000 by default, *-r* parameter), for a maximum number of sequences (100,000 by default, *-x* parameter). By default 1 sample is extracted, but as many as desired can be taken, by using the *-n* parameter.

The sequence file in input can be either be passed to the standard input or as last parameter on the command line. By defult a Fasta is expected, unless the *-q* parameter is passed.

The *-p* parameter specifies the prefix to be used, and if the output files can be gzipped using the *-z* parameter.

mgkit.workflow.sampling_utils.**main**()
Main function

mgkit.workflow.sampling_utils.**sample_command**(*options*)

mgkit.workflow.sampling_utils.**set_parser**()
Sets command line arguments parser

mgkit.workflow.sampling_utils.**set_sample_parser**(*parser*)

mgkit.workflow.snp_parser module

mgkit.workflow.taxon_utils module

The script contains commands used to access functionality related to taxonomy, without the need to write ad-hoc code for functionality that can be part of a workflow. One example is access to the the last common ancestor function contained in the *mgkit.taxon*.

Last Common Ancestor (lca and lca_line)

These commands expose the functionality of `last_common_ancestor_multiple()`, making it accessible via the command line. They differ in the input file format and the choice of output files.

the *lca* command can be used to define the last common ancestor of contigs from the annotation in a GFF file. The command uses the *taxon_ids* from all annotations belonging to a contig/sequence, if they have a **bitscore** higher or equal to the one passed (50 by default). The default output of the command is a tab separated file where the first column is the contig/sequence name, the taxon_id of the last common ancestor, its scientific/common name and its lineage.

For example:

```
contig_21   172788  uncultured phototrophic eukaryote   cellular organisms,
↪environmental samples
```

If the *-r* is used, by passing the fasta file containing the nucleotide sequences the output file is a GFF where for each an annotation for the full contig length contains the same information of the tab separated file format.

The **lca_line** command accept as input a file where each line consist of a list of taxon_ids. The separator for the list can be changed and it defaults to TAB. The last common ancestor for all taxa on a line is searched. The ouput of this command is the same as the tab separated file of the **lca** command, with the difference that instead of the first column, which in this command becames a list of all *taxon_ids* that were used to find the last common ancestor for that line. The list of *taxon_ids* is separated by semicolon ";".

Note: Both also accept the *-n* option, to report the config/line and the taxon_ids that had no common ancestors. These are treated as errors and do not appear in the output file.

Krona Output

New in version 0.3.0.

The *lca* command supports the writing of a file compatible with Krona. The output file can be used with the *ktImportText/ImportText.pl* script included with KronaTools[598]. Specifically, the output from *taxon_utils* will be a file with all the lineages found (tab separated), that can be used with:

```
$ ktImportText -q taxon_utils_ouput
```

Note the use of *-q* to make the script count the lineages. Sequences with no LCA found will be marked as *No LCA* in the graph, the *-n* is not required.

Note: Please note that the output won't include any sequence that didn't have a hit with the software used. If that's important, the **-kt** option can be used to add a number of *Unknown* lines at the end, to read the total supplied.

---

[598] https://github.com/marbl/Krona/wiki

Filter by Taxon

The **filter** command of this script allows to filter a GFF file using the *taxon_id* attribute to include only some annotations, or exclude some. The filter is based on the *mgkit.taxon.is_ancestor* function, and the *mgkit.filter.taxon.filter_taxon_by_id_list*. It allows to pass a list of taxon_id (or taxon_names) to the script. The *include* filter will only output annotations that have one of the passed taxa as ancestors, while the *exclude* filter will remove those annotations, that have the passed taxa as ancestors, from the output.

A list of comma separated taxon_ids can be supplied, as for the names. If any of the the supplied names have multiple taxon_id (e.g. Actinobacteria) the script exits and in the log can be found the list of duplicates. For cases like this, it's preferred for the user to supply a taxon_id, as they can be searched in NCBI taxonomy (also Uniprot).

Warning: Annotations with no taxon_id are not included in the output of both filters

Convert Taxa Tables to HDF5

This command is used to convert the taxa tables download from Uniprot and NCBI, using the scripts mentioned in *download-data - Download Taxonomy from NCBI*, *download-uniprot-taxa.sh* and *download-ncbi-taxa* into a HDF5 file that can be used with the *addtaxa* command in *add-gff-info - Add informations to GFF annotations*.

The advantage is a faster lookup of the IDs. The other is a smaller memory footprint when a great number of annotations are kept in memory.

Changes

Changed in version 0.3.1: added *to_hdf* command

Changed in version 0.3.1: added *-j* option to *lca*, which outputs a JSON file with the LCA results

Changed in version 0.3.0: added *-k* and *-kt* options for Krona output, lineage now includes the LCA also added *-a* option to select between lineages with only ranked taxa. Now it defaults to all components.

Changed in version 0.2.6: added *feat-type* option to *lca* command, added phylum output to nolca

New in version 0.2.5.

mgkit.workflow.taxon_utils.**filter_taxa_command**(*options*)

mgkit.workflow.taxon_utils.**get_taxon_info**(*taxonomy*, *taxon_id*, *only_ranked*)

mgkit.workflow.taxon_utils.**lca_contig_command**(*options*)

mgkit.workflow.taxon_utils.**lca_line_command**(*options*)

mgkit.workflow.taxon_utils.**lca_options**(*parser*)

mgkit.workflow.taxon_utils.**main**()
Main function

mgkit.workflow.taxon_utils.**set_common_options**(*parser*)

mgkit.workflow.taxon_utils.**set_filter_taxa_parser**(*parser*)

mgkit.workflow.taxon_utils.**set_lca_contig_parser**(*parser*)

mgkit.workflow.taxon_utils.**set_lca_line_parser**(*parser*)

mgkit.workflow.taxon_utils.**set_parser**()
Sets command line arguments parser

mgkit.workflow.taxon_utils.**set_taxa_table_parser**(*parser*)

mgkit.workflow.taxon_utils.**taxa_table_command**(*options*)

mgkit.workflow.taxon_utils.**validate_taxon_ids**(*taxon_ids*, *taxonomy*)

mgkit.workflow.taxon_utils.**validate_taxon_names**(*taxon_names*, *taxonomy*)

mgkit.workflow.taxon_utils.**write_json**(*lca_dict*, *seq_id*, *taxonomy*, *taxon_id*, *only_ranked*)

mgkit.workflow.taxon_utils.**write_krona**(*file_handle*, *taxonomy*, *taxon_id*, *only_ranked*)

mgkit.workflow.taxon_utils.**write_lca_gff**(*file_handle*, *seq_id*, *seq*, *taxon_id*, *taxon_name*, *lineage*, *feat_type*)

mgkit.workflow.taxon_utils.**write_lca_tab**(*file_handle*, *seq_id*, *taxon_id*, *taxon_name*, *rank*, *lineage*)

mgkit.workflow.taxon_utils.**write_no_lca**(*file_handle*, *seq_id*, *taxon_ids*, *extra=None*)

mgkit.workflow.utils module

Utility functions for workflows

class mgkit.workflow.utils.**CiteAction**(*option_strings*, *dest='==SUPPRESS=='*, *default='==SUPPRESS=='*, *help='Show citation for the framework'*)

Bases: argparse.Action[599]

Argparse action to print the citation, using the *mgkit.cite()* function.

class mgkit.workflow.utils.**PrintManAction**(*option_strings*, *dest='==SUPPRESS=='*, *default='==SUPPRESS=='*, *help='Show the script manual'*, *manual=''*)

Bases: argparse.Action[600]

New in version 0.2.6.

Argparse action to print the manual

mgkit.workflow.utils.**add_basic_options**(*parser*, *manual=''*)

Changed in version 0.2.6: added *quiet* option

Adds verbose and version options to the option parser

mgkit.workflow.utils.**exit_script**(*message*, *ret_value*)

Used to exit the script with a return value

Module contents

Workflows used to script the library - execute bits of the pipelines supported

7.1.2 Submodules

mgkit.align module

Module dealing with BAM/SAM files

class mgkit.align.**SamtoolsDepth**(*file_handle*, *num_seqs=10000*)

Bases: object[601]

New in version 0.3.0.

[599] https://docs.python.org/3/library/argparse.html#argparse.Action
[600] https://docs.python.org/3/library/argparse.html#argparse.Action
[601] https://docs.python.org/3/library/functions.html#object

A class used to cache the results of *read_samtools_depth()*, while reading only the necessary data from a‘samtools depth -aa‘ file.

data = None

file_handle = None

region_coverage(*seq_id*, *start*, *end*)
Returns the mean coverage of a region. The *start* and *end* parameters are expected to be 1-based coordinates, like the correspondent attributes in *mgkit.io.gff.Annotation* or *mgkit.io.gff.GenomicRange*.

If the sequence for which the coverage is requested is not found, the *depth* file is read (and cached) until it is found.

Parameters

- **seq_id** (*str*[602]) – sequence for which to return mean coverage
- **start** (*int*[603]) – start of the region
- **end** (*int*[604]) – end of the region

Returns mean coverage of the requested region

Return type float[605]

mgkit.align.**add_coverage_info**(*annotations*, *bam_files*, *samples*, *attr_suff='_cov'*)
Adds coverage information to annotations, using BAM files.

The coverage information is added for each sample as a ‘sample_cov’ and the total coverage as as ‘cov’ attribute in the annotations.

Note: The bam_files and sample variables must have the same order

---

Parameters

- **annotations** (*iterable*) – iterable of annotations
- **bam_files** (*iterable*) – iterable of `pysam.Samfile` instances
- **sample** (*iterable*) – names of the samples for the BAM files

mgkit.align.**covered_annotation_bp**(*files*, *annotations*, *min_cov=1*, *progress=False*)
New in version 0.1.14.

Returns the number of base pairs covered of annotations over multiple samples.

Parameters

- **files** (*iterable*) – an iterable that returns the alignment file names
- **annotations** (*iterable*) – an iterable that returns annotations
- **min_cov** (*int*[606]) – minumum coverage for a base to counted
- **progress** (*bool*[607]) – if *True*, a progress bar is used

Returns a dictionary whose keys are the uid and the values the number of bases that are covered by reads among all samples

Return type dict[608]

[602] https://docs.python.org/3/library/stdtypes.html#str
[603] https://docs.python.org/3/library/functions.html#int
[604] https://docs.python.org/3/library/functions.html#int
[605] https://docs.python.org/3/library/functions.html#float
[606] https://docs.python.org/3/library/functions.html#int
[607] https://docs.python.org/3/library/functions.html#bool
[608] https://docs.python.org/3/library/stdtypes.html#dict

mgkit.align.**get_region_coverage**(*bam_file*, *seq_id*, *feat_from*, *feat_to*)
Return coverage for an annotation.

Note: feat_from and feat_to are 1-based indexes

Parameters

- **bam_file** (*Samfile*) – instance of pysam.Samfile
- **seq_id** (*str*[609]) – sequence id
- **feat_from** (*int*[610]) – start position of feature
- **feat_to** (*int*[611]) – end position of feature

Return int coverage array for the annotation

mgkit.align.**read_samtools_depth**(*file_handle*, *num_seqs=10000*)
New in version 0.3.0.

Reads a samtools *depth* file, returning a generator that yields the array of each base coverage on a per-sequence base.

Note: The information on position is not used, to use numpy and save memory. samtools *depth* should be called with the *-aa* option:

```
`samtools depth -aa bamfile`
```

This options will output both base position with 0 coverage and sequneces with no aligned reads

---

Parameters

- **file_handle** (*file*) – file handle of the coverage file
- **num_seqs** (*int*[612]) – number of sequence that fires a log message

Yields *tuple* – the first element is the sequence identifier and the second one is the *numpy* array with the positions

mgkit.consts module

Module containing constants for the filter package

mgkit.graphs module

New in version 0.1.12.

Graph module

mgkit.graphs.**add_module_compounds**(*graph*, *rn_defs*)
New in version 0.3.1.

Modify in-place a graph, by adding additional compounds from a dictionary of definitions. It uses the reversible/irreversible information for each reaction to add the correct number of edges to the graph.

Parameters

[609] https://docs.python.org/3/library/stdtypes.html#str
[610] https://docs.python.org/3/library/functions.html#int
[611] https://docs.python.org/3/library/functions.html#int
[612] https://docs.python.org/3/library/functions.html#int

- **graph** (*graph*) – a graph to update with additional compounds
- **rn_defs** (*dict*[613]) – a dictionary, whose keys are reactions IDs and the values are instances of *mgkit.kegg.KeggReaction*

mgkit.graphs.**annotate_graph_nodes**(*graph*, *attr*, *id_map*, *default=None*)

New in version 0.1.12.

Add/Changes nodes attribute *attr* using a dictionary of ids->values.

Note: If the id is not found in *id_map*:

- default is None: no value added for that node
- default is not None: the node attribute will be set to *default*

Parameters

- **graph** – the graph to annotate
- **attr** (*str*[614]) – the attribute to annotate
- **id_map** (*dict*[615]) – the dictionary with the values for each node
- **default** – the value used in case an *id* is not found in *id_map*

mgkit.graphs.**build_graph**(*id_links*, *name*, *edge_type=''*, *weight=0.5*)

New in version 0.1.12.

Builds a networkx graph from a dictionary of nodes, as outputted by *mgkit.kegg.KeggClientRest.get_pathway_links()*. The graph is undirected, and all edges weight are the same.

Parameters

- **id_links** (*dict*[616]) – dictionary with the links
- **name** (*str*[617]) – name of the graph
- **edge_type** (*str*[618]) – an optional name for the *edge_type* attribute set for each edge
- **weight** (*float*[619]) – the weight assigned to each edge in the graph

Returns an instance of `networkx.Graph`

Return type graph

mgkit.graphs.**build_weighted_graph**(*id_links*, *name*, *weights*, *edge_type=''*)

New in version 0.1.14.

Builds a networkx graph from a dictionary of nodes, as outputted by *mgkit.kegg.KeggClientRest.get_pathway_links()*. The graph is undirected, and all edges weight are the same.

Parameters

- **id_links** (*dict*[620]) – dictionary with the links
- **name** (*str*[621]) – name of the graph
- **edge_type** (*str*[622]) – an optional name for the *edge_type* attribute set for each edge

[613] https://docs.python.org/3/library/stdtypes.html#dict
[614] https://docs.python.org/3/library/stdtypes.html#str
[615] https://docs.python.org/3/library/stdtypes.html#dict
[616] https://docs.python.org/3/library/stdtypes.html#dict
[617] https://docs.python.org/3/library/stdtypes.html#str
[618] https://docs.python.org/3/library/stdtypes.html#str
[619] https://docs.python.org/3/library/functions.html#float
[620] https://docs.python.org/3/library/stdtypes.html#dict
[621] https://docs.python.org/3/library/stdtypes.html#str
[622] https://docs.python.org/3/library/stdtypes.html#str

- **weight** (*float*[623]) – the weight assigned to each edge in the graph

Returns an instance of `networkx.Graph`

Return type graph

`mgkit.graphs.`**copy_edges**(*g*, *graph1*, *name=None*, ***kwd*)
New in version 0.1.12.

Used by *link_nodes()* to copy edges

`mgkit.graphs.`**copy_nodes**(*g*, *graph1*, *name=None*, *id_attr=None*, ***kwd*)
New in version 0.1.12.

Used by *link_nodes()* to copy nodes

`mgkit.graphs.`**filter_graph**(*graph*, *id_list*, *filter_func=<function <lambda>>*)
New in version 0.1.12.

Filter a graph based on the *id_list* provided and the filter function used to test the id attribute of each node.

A node is removed if *filter_func* returns True on a node and its id attribute is not in *id_list*

Parameters

- **graph** – the graph to filter
- **id_list** (*iterable*) – the list of nodes that are to remain in the graph
- **filter_func** (*func*) – function which accept a single parameter and return a boolean

Returns an instance of `networkx.Graph`

Return type graph

`mgkit.graphs.`**from_kgml**(*entry*, *graph=None*, *rn_ids=None*)
New in version 0.3.1.

Given a KGML file (as string), representing a pathway in Kegg, returns a networkx DiGraph, using reaction directionality included in the KGML. If a reaction is reversible, 2 edges (from and to) for each compound/reaction pair are added, giving the bidirectionality.

Note: substrate and products included in a KGML don't represent the complete reaction, excluding in general cofactors or more general terms. Those can be added using *add_module_compounds()*, which may be more useful when used with a restricted number of reactions (e.g. a module)

Parameters

- **entry** (*str*[624]) – KGML file as a string, or anything that can be passed to ElementTree
- **graph** (*graph*) – an instance of a networkx DiGraph if the network is to be updated with a new KGML, if *None* a new one is created
- **rn_ids** (*set*[625]) – a set/list of reaction IDs that are to be included, if *None* all reactions are used

Returns a networkx DiGraph with the reaction/compounds

Return type graph

`mgkit.graphs.`**link_graph**(*graphs*, *edge_links*)
New in version 0.1.12.

[623] https://docs.python.org/3/library/functions.html#float
[624] https://docs.python.org/3/library/stdtypes.html#str
[625] https://docs.python.org/3/library/stdtypes.html#set

Link nodes of a set of graphs using the specifics in edge_links. The resulting graph nodes are renamed, and the nodes that are shared between the graphs linked.

Parameters

- **graphs** – iterable of graphs
- **edge_links** – iterable with function, edge_type and weight for the links between graphs

Returns an instance of `networkx.Graph`

Return type graph

`mgkit.graphs.`**link_nodes**(*g*, *graph1*, *graph2*, *id_filter*, *link_type*, *weight*)

New in version 0.1.12.

Used by *link_graph()* to link nodes with the same *id*

`mgkit.graphs.`**rename_graph_nodes**(*graph*, *name_func=None*, *exclude_ids=None*)

mgkit.kegg module

Module containing classes and functions to access Kegg data

class `mgkit.kegg.`**KeggClientRest**(*cache=None*)

Bases: `object`[626]

Changed in version 0.3.1: added a *cache* attribute for some methods

Kegg REST client

The class includes methods and data to use the REST API provided by Kegg. At the moment it provides methods to for 'link', 'list' and 'get' operations,

Kegg REST API[627]

api_url = 'http://rest.kegg.jp/'

cache = None

contact = None

conv(*target_db*, *source_db*, *strip=True*)

New in version 0.3.1.

Kegg Help:

http://rest.kegg.jp/conv/<target_db>/<source_db>

(<target_db> <source_db>) = (<kegg_db> <outside_db>) | (<outside_db> <kegg_db>)

For gene identifiers: <kegg_db> = <org> <org> = KEGG organism code or T number <outside_db> = ncbi-proteinid | ncbi-geneid | uniprot

For chemical substance identifiers: <kegg_db> = drug | compound | glycan <outside_db> = pubchem | chebi http://rest.kegg.jp/conv/<target_db>/<dbentries>

For gene identifiers: <dbentries> = database entries involving the following <database> <database> = <org> | genes | ncbi-proteinid | ncbi-geneid | uniprot <org> = KEGG organism code or T number

For chemical substance identifiers: <dbentries> = database entries involving the following <database> <database> = drug | compound | glycan | pubchem | chebi

[626] https://docs.python.org/3/library/functions.html#object
[627] http://www.kegg.jp/kegg/rest/keggapi.html

Examples

```
>>> kc = KeggClientRest()
>>> kc.conv('ncbi-geneid', 'eco')
{'eco:b0217': {'ncbi-geneid:949009'},
 'eco:b0216': {'ncbi-geneid:947541'},
 'eco:b0215': {'ncbi-geneid:946441'},
 'eco:b0214': {'ncbi-geneid:946955'},
 'eco:b0213': {'ncbi-geneid:944903'},
...
>>> kc.conv('ncbi-proteinid', 'hsa:10458+ece:Z5100')
{'10458': {'NP_059345'}, 'Z5100': {'AAG58814'}}
```

cpd_desc_re = <_sre.SRE_Pattern object>

cpd_re = <_sre.SRE_Pattern object at 0x28389e0>

empty_cache(*methods=None*)

New in version 0.3.1.

Empties the cache completely or for a specific method(s)

Parameters **methods** (*iterable, str*[628]) – string or iterable of strings that are part of the cache. If None the cache is fully emptied

find(*query*, *database*, *options=None*, *strip=True*)

New in version 0.3.1.

Kegg Help:

http://rest.kegg.jp/find/<database>/<query>

<database> = pathway | module | ko | genome | <org> | compound | glycan | reaction | rclass | enzyme | disease | drug | dgroup | environ | genes | ligand

<org> = KEGG organism code or T number

http://rest.kegg.jp/find/<database>/<query>/<option>

<database> = compound | drug <option> = formula | exact_mass | mol_weight

Examples

```
>>> kc = KeggClientRest()
>>> kc.find('CH4', 'compound')
{'C01438': 'Methane; CH4'}
>>> kc.find('K00844', 'genes', strip=False)
{'tped:TPE_0072': 'hexokinase; K00844 hexokinase [EC:2.7.1.1]',
...
>>> kc.find('174.05', 'compound', options='exact_mass')
{'C00493': '174.052823',
 'C04236': '174.052823',
 'C16588': '174.052823',
 'C17696': '174.052823',
 'C18307': '174.052823',
 'C18312': '174.052823',
 'C21281': '174.052823'}
```

get_entry(*k_id*, *option=None*)

Changed in version 0.3.1: this is now cached

The method abstract the use of the 'get' operation in the Kegg API

[628] https://docs.python.org/3/library/stdtypes.html#str

Parameters

- **k_id** (*str*[629]) – kegg id of the resource to get
- **option** (*str*[630]) – optional, to specify a format

get_ids_names(*target='ko'*, *strip=True*)
New in version 0.1.13.

Changed in version 0.3.1: the call is now cached

Returns a dictionary with the names/description of all the id of a specific target, (ko, path, cpd, etc.)

If strip=True the id will stripped of the module abbreviation (e.g. md:M00002->M00002)

get_ortholog_pathways()
Gets ortholog pathways, replace 'map' with 'ko' in the id

get_pathway_links(*pathway*)
Returns a dictionary with the mappings KO->compounds for a specific Pathway or module

get_reaction_equations(*ids*, *max_len=10*)
Get the equation for the reactions

id_prefix = {'C': 'cpd', 'K': 'ko', 'R': 'rn', 'k': 'map', 'm': 'path'}

ko_desc_re = <_sre.SRE_Pattern object>

link(*target*, *source*, *options=None*)
New in version 0.2.0.

Implements "link" operation in Kegg REST

http://www.genome.jp/linkdb/

link_ids(*target*, *kegg_ids*, *max_len=50*)
Changed in version 0.3.1: removed *strip* and cached the results

The method abstract the use of the 'link' operation in the Kegg API

The target parameter can be one of the following:

```
pathway | brite | module | disease | drug | environ | ko | genome |
<org> | compound | glycan | reaction | rpair | rclass | enzyme

<org> = KEGG organism code or T number
```

Parameters

- **target** (*str*[631]) – the target db
- **ids** – can be either a single id as a string or a list of ids
- **strip** (*bool*[632]) – if the prefix (e.g. ko:K00601) should be stripped
- **max_len** (*int*[633]) – the maximum number of ids to retrieve with each request, should not exceed 50

Return dict dictionary mapping requested id to target id(s)

list_ids(*k_id*)
The method abstract the use of the 'list' operation in the Kegg API

The k_id parameter can be one of the following:

[629] https://docs.python.org/3/library/stdtypes.html#str
[630] https://docs.python.org/3/library/stdtypes.html#str
[631] https://docs.python.org/3/library/stdtypes.html#str
[632] https://docs.python.org/3/library/functions.html#bool
[633] https://docs.python.org/3/library/functions.html#int

```
pathway | brite | module | disease | drug | environ | ko | genome |
<org> | compound | glycan | reaction | rpair | rclass | enzyme

<org> = KEGG organism code or T number
```

Parameters **k_id** (*str*[634]) – kegg database to get list of ids

Return list list of ids in the specified database

load_cache(*file_handle*)

New in version 0.3.1.

Loads the cache from file

rn_eq_re = <_sre.SRE_Pattern object>

rn_name_re = <_sre.SRE_Pattern object>

write_cache(*file_handle*)

New in version 0.3.1.

Write the cache to file

class mgkit.kegg.**KeggCompound**(*cp_id=None*, *description=''*)

Bases: object[635]

Kegg compound

__eq__(*other*)

```
>>> KeggCompound('test') == KeggCompound('test')
True
>>> KeggCompound('test') == 1
False
```

__ne__(*other*)

```
>>> KeggCompound('test') != KeggCompound('test1')
True
>>> KeggCompound('test') != 1
True
```

class mgkit.kegg.**KeggData**(*fname=None*, *gen_maps=True*)

Bases: object[636]

gen_ko_map()

gen_maps()

get_cp_names()

get_ko_cp_links(*path_filter=None*, *description=False*)

get_ko_cp_links_alt(*direction='out'*, *description=False*)

get_ko_names()

get_ko_pathway_map(*black_list=None*)

get_ko_pathways(*ko_id*)

get_ko_rn_links(*path_filter=None*, *description=False*)

[634] https://docs.python.org/3/library/stdtypes.html#str
[635] https://docs.python.org/3/library/functions.html#object
[636] https://docs.python.org/3/library/functions.html#object

get_pathway_ko_map(*black_list=None*)

get_rn_cp_links(*path_filter=None*, *description=False*)

get_rn_names()

load_data(*fname*)

maps = None

pathways = None

save_data(*fname*)

class mgkit.kegg.**KeggMapperBase**(*fname=None*)
Bases: object[637]

Base object for Kegg mapping classes

get_id_map()
Returns a mapping->KOs dictionary (a reverse mapping to get_ko_map)

get_id_names()
Returns a copy of the mapping names

get_ko_map()
Returns a copy of the KO->mapping dictionary

static ko_to_mapping(*ko_id*, *query*, *columns*, *contact=None*)
Returns the mappings to the supplied KO. Can be used for any id, the query format is free as well as the columns returned. The only restriction is using a tab format, that is parsed.

Parameters

- **ko_id** (*str*[638]) – id used in the query
- **query** (*str*[639]) – query passed to the Uniprot API, ko_id is replaced using `str.format()`
- **column** (*str*[640]) – column used in the results table used to map the ids
- **contact** (*str*[641]) – email address to be passed in the query (requested Uniprot API)

Note: each mapping in the column is separated by a ;

load_data(*fname*)
Loads mapping data to disk

save_data(*fname*)
Saves mapping data to disk

class mgkit.kegg.**KeggModule**(*entry=None*, *old=False*)
Bases: object[642]

New in version 0.1.13.

Used to extract information from a pathway module entry in Kegg

The entry, as a string, can be either passed at instance creation or with *KeggModule.parse_entry()*

classes = None

compounds = None

[637] https://docs.python.org/3/library/functions.html#object
[638] https://docs.python.org/3/library/stdtypes.html#str
[639] https://docs.python.org/3/library/stdtypes.html#str
[640] https://docs.python.org/3/library/stdtypes.html#str
[641] https://docs.python.org/3/library/stdtypes.html#str
[642] https://docs.python.org/3/library/functions.html#object

entry = ''

find_submodules()
New in version 0.3.0.

Returns the possible submodules, as a list of tuples where the elements are the first and last compounds in a submodule

first_cp
Returns the first compound in the module

last_cp
Returns the first compound in the module

name = ''

parse_entry(*entry*)
Parses a Kegg module entry and change the instance values. By default the reactions IDs are substituted with the KO IDs

parse_entry2(*entry*)
New in version 0.3.0.

Parses a Kegg module entry and change the instance values. By default the reactions IDs are NOT substituted with the KO IDs.

static parse_reaction(*line*, *ko_ids=None*)
Changed in version 0.3.0: cleaned the parsing

parses the lines with the reactions and substitute reaction IDs with the corresponding KO IDs if provided

reactions = None

to_edges(*id_only=None*)
Changed in version 0.3.0: added id_only and changed to reflect changes in *reactions*

Returns the reactions as edges that can be supplied to make graph.

Parameters **id_only** (*None, iterable*) – if None the returned edges are for the whole module, if an iterable (converted to a set[643]), only edges for those reactions are returned

Yields *tuple* – the elements are the compounds and reactions in the module

class mgkit.kegg.**KeggOrtholog**(*ko_id=None*, *description=''*, *reactions=None*)
Bases: object[644]

Kegg Ortholog gene

__eq__(*other*)

```
>>> KeggOrtholog('test') == KeggOrtholog('test')
True
>>> KeggOrtholog('test') == 1
False
```

__ne__(*other*)

```
>>> KeggOrtholog('test') != KeggOrtholog('test1')
True
>>> KeggOrtholog('test') != 1
True
```

[643] https://docs.python.org/3/library/stdtypes.html#set
[644] https://docs.python.org/3/library/functions.html#object

class mgkit.kegg.**KeggPathway**(*path_id=None*, *description=None*, *genes=None*)
Bases: object[645]

Kegg Pathway

__eq__(*other*)

```
>>> KeggPathway('test') == KeggPathway('test')
True
>>> KeggPathway('test') == 1
False
```

__ne__(*other*)

```
>>> KeggPathway('test') != KeggPathway('test1')
True
>>> KeggPathway('test') != 1
True
```

class mgkit.kegg.**KeggReaction**(*entry*)
Bases: object[646]

Changed in version 0.3.1: reworked, only stores the equation

Kegg Reaction, used for parsing the equation line

left_cp = None

right_cp = None

rn_id = None

mgkit.kegg.**download_data**(*fname='kegg.pickle'*, *contact=None*)

mgkit.kegg.**parse_reaction**(*line*, *prefix=('C', 'G')*)
New in version 0.3.1.

Parses a reaction equation from Kegg, returning the left and right components. Needs testing

Parameters **line** (*str*[647]) – reaction string

Returns left and right components as *sets*

Return type tuple[648]

Raises ValueError[649] – if the

mgkit.logger module

Module configuring log information

mgkit.logger.**config_log**(*level=10*, *output=<open file '<stderr>', mode 'w'>*)
Minimal configuration of :mod‘logging‘ module, default to debug level and the output is printed to standard error

Parameters

- **level** (*int*[650]) – logging level
- **output** (*file*) – file to which write the log

[645] https://docs.python.org/3/library/functions.html#object
[646] https://docs.python.org/3/library/functions.html#object
[647] https://docs.python.org/3/library/stdtypes.html#str
[648] https://docs.python.org/3/library/stdtypes.html#tuple
[649] https://docs.python.org/3/library/exceptions.html#ValueError
[650] https://docs.python.org/3/library/functions.html#int

mgkit.logger.**config_log_to_file**(*level=10*, *output=None*)

New in version 0.1.14.

Minimal configuration of :mod‘logging‘ module, default to debug level and the output is printed to script name, using *sys.argv[0]*.

Parameters

- **level** (*int*[651]) – logging level
- **output** (*file*) – file to which write the log

mgkit.simple_cache module

class mgkit.simple_cache.**memoize**(*func*)

Bases: dict[652]

a cache found on the PythonDecoratorLibrary[653]

Not sure about the license for it.

mgkit.taxon module

This module gives access to Uniprot taxonomy data. It also defines classes to filter, order and group data by taxa

exception mgkit.taxon.**NoLcaFound**

Bases: exceptions.Exception

New in version 0.1.13.

Raised if no lowest common ancestor can be found in the taxonomy

class mgkit.taxon.**UniprotTaxonTuple**(*taxon_id*, *s_name*, *c_name*, *rank*, *lineage*, *parent_id*)

Bases: tuple[654]

__getnewargs__()

Return self as a plain tuple. Used by copy and pickle.

__getstate__()

Exclude the OrderedDict from pickling

__repr__()

Return a nicely formatted representation string

_asdict()

Return a new OrderedDict which maps field names to their values

_replace(***kwds*)

Return a new UniprotTaxonTuple object replacing specified fields with new values

c_name

Alias for field number 2

lineage

Alias for field number 4

parent_id

Alias for field number 5

rank

Alias for field number 3

[651] https://docs.python.org/3/library/functions.html#int
[652] https://docs.python.org/3/library/stdtypes.html#dict
[653] https://wiki.python.org/moin/PythonDecoratorLibrary#Alternate_memoize_as_dict_subclass
[654] https://docs.python.org/3/library/stdtypes.html#tuple

s_name
Alias for field number 1

taxon_id
Alias for field number 0

class mgkit.taxon.**UniprotTaxonomy**(*fname=None*)
Bases: object[655]

Class that contains the whole Uniprot taxonomy. Defines some methods to easy access of taxonomy.

Defines:

- methods to load taxonomy from a pickle file or a generic file handle
- can be iterated over and returns a generator its UniprotTaxon instances
- can be used as a dictionary, in which the key is a taxon_id and the value is its UniprotTaxon instance

__contains__(*taxon*)
Returns True if the taxon is in the taxonomy

Accepts an int (check for taxon_id) or an instance of UniprotTaxon

__getitem__(*taxon_id*)
Defines dictionary behavior. Key is a taxon_id, the returned value is a UniprotTaxon instance

__iter__()
Defines iterable behavior. Returns a generator for UniprotTaxon instances

__len__()
Returns the number of taxa contained

__repr__()
New in version 0.2.5.

add_lineage(***lineage*)
New in version 0.3.1.

Adds a lineage to the taxonomy. It's passed by keyword arguments, where each key is a value in the *TAXON_RANKS* rankes and the value is the scientific name. Appended underscores '_' will be stripped from the rank name. This is for cases such as *class* where the key is a reserved word in Python. Also one extra node can be added, such as strain/cultivar/subspecies and so on, but one only is expected to be passed.

Parameters **lineage** (*dict*[656]) – the lineage as a keyword arguments

Returns the taxon_id of the last element in the lineage

Return type int[657]

Raises

- ValueError[658] – if more than a keyword argument is not contained in
- *TAXON_RANKS*

add_taxon(*taxon_name*, *common_name=''*, *rank='no rank'*, *parent_id=None*)
New in version 0.3.1.

Adds a taxon to the taxonomy. If a taxon with the same name and rank is found, its taxon_id is returned, otherwise a new taxon_id is returned.

Parameters

- **taxon_name** (*str*[659]) – scientific name of the taxon

[655] https://docs.python.org/3/library/functions.html#object
[656] https://docs.python.org/3/library/stdtypes.html#dict
[657] https://docs.python.org/3/library/functions.html#int
[658] https://docs.python.org/3/library/exceptions.html#ValueError
[659] https://docs.python.org/3/library/stdtypes.html#str

- **common_name** (*str*[660]) – common name
- **rank** (*str*[661]) – rank, defaults to 'no rank'
- **parent_id** (*int*[662]) – taxon_id of the parent, defaults to *None*, which is the taxonomy root

Returns the taxon_id of the added taxon (if new), or the taxon_id of the taxon with the same name and rank found in the taxonomy

Return type int[663]

Raises

- KeyError[664] – if more than one taxon has already the passed name and
- rank and it can't be resolved by looking at the parent_id passed,
- the exception is raised.

drop_taxon(*taxon_id*)

New in version 0.3.1.

Drops a taxon and all taxa below it in the taxonomy. Also reset the name map for conistency.

Parameters **taxon_id** (*int*[665]) – taxon_id to drop from the taxonomy

find_by_name(*s_name*, *rank=None*, *strict=True*)

Changed in version 0.2.3: the search is now case insensitive

Changed in version 0.3.1: added *rank* and *strict* parameter

Returns the taxon IDs associated with the scientific name provided

Parameters

- **s_name** (*str*[666]) – the scientific name
- **rank** (*str*[667], *None*) – return only a taxon_id of a specific rank
- **strict** (*book*) – if True and *rank* is not None, KeyError will be raised if multiple taxa have the same name and rank

Returns a reference to the list of IDs that have for *s_name*, if *rank* is None. If *rank* is not None and one taxon is found, its taxon_id is returned, or None if no taxon is found. If *strict* is True and *rank* is not None, the set of taxon_ids found is resturned.

Return type list[668]

Raises

- KeyError[669] – If multiple taxa are found, a *KeyError* exception is
- raised.

gen_name_map()

Changed in version 0.2.3: names are stored in the mapping as lowercase

Generate a name map, where to each scientific name in the taxonomy an id is associated.

[660] https://docs.python.org/3/library/stdtypes.html#str
[661] https://docs.python.org/3/library/stdtypes.html#str
[662] https://docs.python.org/3/library/functions.html#int
[663] https://docs.python.org/3/library/functions.html#int
[664] https://docs.python.org/3/library/exceptions.html#KeyError
[665] https://docs.python.org/3/library/functions.html#int
[666] https://docs.python.org/3/library/stdtypes.html#str
[667] https://docs.python.org/3/library/stdtypes.html#str
[668] https://docs.python.org/3/library/stdtypes.html#list
[669] https://docs.python.org/3/library/exceptions.html#KeyError

get_lineage(*taxon_id*, *names=False*, *only_ranked=True*, *with_last=True*)
New in version 0.3.1.

Proxy for *get_lineage()*, with changed defaults

get_name_map()
Returns a taxon_id->s_name dictionary

get_ranked_taxon(*taxon_id*, *rank=None*, *ranks=('superkingdom', 'kingdom', 'phylum', 'class', 'subclass', 'order', 'family', 'genus', 'species')*, *roots=False*)
Changed in version 0.1.13: added *roots* argument

Traverse the branch of which the *taxon* argument is the leaf backward, to get the specific rank to which the *taxon* belongs to.

> **Warning:** the *roots* options is kept for backward compatibility and should be be set to *False*

Parameters

- **taxon_id** – id of the taxon or instance of `UniprotTaxon`
- **rank** (*str*[670]) – string that specify the rank, if None, the first valid rank will be searched. (i.e. the first with a value different from '')
- **ranks** – tuple of all taxonomy ranks, default to the default module value
- **roots** (*bool*[671]) – if True, uses `TAXON_ROOTS` to solve the root taxa

Returns instance of `UniprotTaxon` for the rank found.

is_ancestor(*leaf_id*, *anc_ids*)
Changed in version 0.1.13: now uses *is_ancestor()* and changed behavior

Checks if a taxon is the leaf of another one, or a list of taxa.

Parameters

- **leaf_id** (*int*[672]) – leaf taxon id
- **anc_ids** (*int*[673]) – ancestor taxon id(s)

Return bool True if the ancestor taxon is in the leaf taxon lineage

load_data(*file_handle*)
Changed in version 0.2.3: now can use read *msgpack* serialised files

Changed in version 0.1.13: now accepts file handles and compressed files (if file names)

Loads serialised data from file name "file_handle" and accepts compressed files.

if the *.msgpack* string is found in the file name, the *msgpack* package is used instead of pickle

Parameters **file_handle** (*str*[674], *file*) – file name to which save the instance data

Raises

- `DependencyError` – if the file name contains *.msgpack* and the
- package is not installed

read_from_gtdb_taxonomy(*file_handle*, *use_gtdb_name=True*, *sep='\t'*)
New in version 0.3.0.

Changed in version 0.3.1: replaced *domain* with *superkingdom* to support *get_lineage*

[670] https://docs.python.org/3/library/stdtypes.html#str
[671] https://docs.python.org/3/library/functions.html#bool
[672] https://docs.python.org/3/library/functions.html#int
[673] https://docs.python.org/3/library/functions.html#int
[674] https://docs.python.org/3/library/stdtypes.html#str

Reads a GTDB taxonomy file (tab separated genome_id/taxonomy) and populate the taxonomy instance. The method also return a dictionary of genome_id -> taxon_id.

Parameters

- **file_handle** (*file*) – file with the taxonomy
- **use_gtdb_name** (*bool*[675]) – if True, the names are kept as-is in the *s_name* attribute of *UniprotTaxonTuple* and the "cleaned" version in *c_name* (e.g. f__Ammonifexaceae -> Ammonifexaceae). If False, the values are switched
- **sep** (*str*[676]) – separator between the columns of the file

Returns dictionary of genome_id -> taxon_id, reflecting the created taxonomy

Return type dict[677]

Note: the taxon_id are generated, so there's no guarantee they will be the same in a successive execution

---

read_from_ncbi_dump(*nodes_file*, *names_file=None*, *merged_file=None*)
New in version 0.2.3.

Uses the *nodes.dmp* and optionally *names.dmp*, *merged.dmp* files from ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/ to populate the taxonomy.

Parameters

- **nodes_file** (*str*[678], *file*) – file name or handle to the file
- **names_file** (*str*[679], *file*, *None*) – file name or handle to the file, if None, names won't be added to the taxa
- **merged_file** (*str*[680], *file*, *None*) – file name or handle to the file, if None, pointers to merged taxa won't be added

read_taxonomy(*f_handle*, *light=True*)
Changed in version 0.2.1: added *light* parameter

Reads taxonomy from a file handle. The file needs to be a tab separated format return by a query on Uniprot. If *light* is True, lineage is not stored to decrease the memory usage. This is now the default.

New taxa will be added, duplicated taxa will be skipped.

Parameters **f_handle** (*handle*) – file handle of the taxonomy file.

save_data(*file_handle*)
Changed in version 0.2.3: now can use *msgpack* to serialise

Saves taxonomy data to a file handle or file name, can write compressed data if the file ends with ".gz", ".bz2"

if the *.msgpack* string is found in the file name, the *msgpack* package is used instead of pickle

Parameters **file_handle** (*str*[681], *file*) – file name to which save the instance data

Raises

- DependencyError – if the file name contains *.msgpack* and the
- package is not installed

675 https://docs.python.org/3/library/functions.html#bool
676 https://docs.python.org/3/library/stdtypes.html#str
677 https://docs.python.org/3/library/stdtypes.html#dict
678 https://docs.python.org/3/library/stdtypes.html#str
679 https://docs.python.org/3/library/stdtypes.html#str
680 https://docs.python.org/3/library/stdtypes.html#str
681 https://docs.python.org/3/library/stdtypes.html#str

mgkit.taxon.**distance_taxa_ancestor**(*taxonomy*, *taxon_id*, *anc_id*)

New in version 0.1.16.

Function to calculate the distance between a taxon and the given ancestor

The distance is equal to the number of step in the taxonomy taken to arrive at the ancestor.

Parameters

- **taxonomy** – *UniprotTaxonomy* instance
- **taxon_id** (*int*[682]) – taxonomic identifier
- **anc_id** (*int*[683]) – taxonomic identifier of the ancestor

Raturns: int: distance between *taxon_id* and it ancestor *anc_id*

mgkit.taxon.**distance_two_taxa**(*taxonomy*, *taxon_id1*, *taxon_id2*)

New in version 0.1.16.

Calculate the distance between two taxa. The distance is equal to the sum steps it takes to traverse the taxonomy until their last common ancestor.

Parameters

- **taxonomy** – *UniprotTaxonomy* instance
- **taxon_id1** (*int*[684]) – taxonomic identifier of first taxon
- **taxon_id2** (*int*[685]) – taxonomic identifier of second taxon

Raturns: int: distance between *taxon_id1* and *taxon_id2*

mgkit.taxon.**get_ancestor_map**(*leaf_ids*, *anc_ids*, *taxonomy*)

This function returns a dictionary where every leaf taxon is associated with the right ancestors in anc_ids

ex. {clostridium: [bacteria, clostridia]}

mgkit.taxon.**get_lineage**(*taxonomy*, *taxon_id*, *names=False*, *only_ranked=False*, *with_last=False*)

New in version 0.2.1.

Changed in version 0.2.5: added *only_ranked*

Changed in version 0.3.0: added *with_last*

Returns the lineage of a taxon_id, as a list of taxon_id or taxa names

Parameters

- **taxonomy** – a *UniprotTaxonomy* instance
- **taxon_id** (*int*[686]) – taxon_id whose lineage to return
- **names** (*bool*[687]) – if True, the returned list contains the names of the taxa instead of the taxon_id
- **only_ranked** (*bool*[688]) – if True, only taxonomic levels whose rank is in data:*TAXON_RANKS* will be returned
- **with_last** (*bool*[689]) – if True, the passed taxon_id is included in the lineage

[682] https://docs.python.org/3/library/functions.html#int
[683] https://docs.python.org/3/library/functions.html#int
[684] https://docs.python.org/3/library/functions.html#int
[685] https://docs.python.org/3/library/functions.html#int
[686] https://docs.python.org/3/library/functions.html#int
[687] https://docs.python.org/3/library/functions.html#bool
[688] https://docs.python.org/3/library/functions.html#bool
[689] https://docs.python.org/3/library/functions.html#bool

Returns lineage of the taxon_id, the elements are *int* if names is False, and *str* when *names* is True. If a taxon has no scientific name, the common name is used. If *only_ranked* is True, the returned list only contains ranked taxa (according to TAXON_RANKS).

Return type list[690]

mgkit.taxon.**group_by_root**(*taxa*, *roots=('archaea', 'bacteria', 'fungi', 'metazoa', 'environmental samples', 'viruses', 'viroidseukaryota', 'other sequences', 'unidentified', 'apusozoa', 'alveolata', 'rhizaria', 'amoebozoa', 'centroheliozoa', 'breviatea', 'oxymonadida', 'parabasalia', 'choanoflagellida', 'jakobida', 'malawimonadidae', 'diplomonadida', 'formicata', 'heterolobosea', 'euglenozoa', 'streptophyta', 'stramenopiles', 'haptophyceae', 'rhodophyta', 'cryptophyta', 'chlorophyta')*, *only_names=False*, *replace_space='#'*)

Deprecated since version 0.2.6.

Returns a dictionary containing as keys the root taxa and as values a list of taxa belonging to that group (checks lineage attribute of UniprotTaxon instances)

Parameters

- **taxa** (*iterable*) – iterable of UniprotTaxon instances
- **roots** – root taxa to be used to construct the dictionary, defaults to the 5 defined in TAXON_ROOTS
- **only_names** (*bool*[691]) – boolean that specify what data type to return for values, if True returns strings (taxa names), if False UniprotTaxon instances
- **replace_space** (*str*[692]) – if only_names is True replaces spaces with the choosen character

Returns a dictionary root->[taxa_ids]

mgkit.taxon.**is_ancestor**(*taxonomy*, *taxon_id*, *anc_id*)

Changed in version 0.1.16: if a *taxon_id* raises a KeyError, False is returned

Determine if the given taxon id (taxon_id) has anc_id as ancestor.

:param *UniprotTaxonomy* taxonomy: taxonomy used to test :param int taxon_id: leaf taxon to test :param int anc_id: ancestor taxon to test against

Return bool True if anc_id is an ancestor of taxon_id or their the same

mgkit.taxon.**last_common_ancestor**(*taxonomy*, *taxon_id1*, *taxon_id2*)

New in version 0.1.13.

Finds the last common ancestor of two taxon IDs. An alias to this function is in the same module, called *lowest_common_ancestor* for compatibility.

Parameters

- **taxonomy** – *UniprotTaxonomy* instance used to test
- **taxon_id1** (*int*[693]) – first taxon ID
- **taxon_id2** (*int*[694]) – second taxon ID

Raturns: int: taxon ID of the lowest common ancestor

Raises *NoLcaFound* – if no common ancestor can be found

[690] https://docs.python.org/3/library/stdtypes.html#list
[691] https://docs.python.org/3/library/functions.html#bool
[692] https://docs.python.org/3/library/stdtypes.html#str
[693] https://docs.python.org/3/library/functions.html#int
[694] https://docs.python.org/3/library/functions.html#int

mgkit.taxon.**last_common_ancestor_multiple**(*taxonomy*, *taxon_ids*)
New in version 0.2.5.

Applies *last_common_ancestor()* to an iterable that yields *taxon_id* while removing any *None* values. If the list is of one element, that *taxon_id* is returned.

Parameters

- **taxonomy** – instance of *UniprotTaxonomy*
- **taxon_ids** (*iterable*) – an iterable that yields taxon_id

Returns the *taxon_id* that is the last common ancestor of all taxon_ids passed

Return type int[695]

Raises

- *NoLcaFound* – when no common ancestry is found or the number of
- *taxon_ids* is 0

mgkit.taxon.**load_taxonomy_map**(*taxon_file*)
Loads taxonomy from file and return a map root_taxon->taxa

mgkit.taxon.**lowest_common_ancestor**(*taxonomy*, *taxon_id1*, *taxon_id2*)
New in version 0.1.13.

Finds the last common ancestor of two taxon IDs. An alias to this function is in the same module, called *lowest_common_ancestor* for compatibility.

Parameters

- **taxonomy** – *UniprotTaxonomy* instance used to test
- **taxon_id1** (*int*[696]) – first taxon ID
- **taxon_id2** (*int*[697]) – second taxon ID

Raturns: int: taxon ID of the lowest common ancestor

Raises *NoLcaFound* – if no common ancestor can be found

mgkit.taxon.**parse_ncbi_taxonomy_merged_file**(*file_handle*)
New in version 0.2.3.

Parses the *merged.dmp* file where the merged taxon_id are stored. Available at ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/

Parameters **file_handle** (*str*[698], *file*) – file name or handle to the file

Returns dictionary with merged_id -> taxon_id

Return type dict[699]

mgkit.taxon.**parse_ncbi_taxonomy_names_file**(*file_handle*, *name_classes=('scientific name', 'common name')*)
New in version 0.2.3.

Parses the *names.dmp* file where the names associated to a taxon_id are stored. Available at ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/

Parameters

- **file_handle** (*str*[700], *file*) – file name or handle to the file

[695] https://docs.python.org/3/library/functions.html#int
[696] https://docs.python.org/3/library/functions.html#int
[697] https://docs.python.org/3/library/functions.html#int
[698] https://docs.python.org/3/library/stdtypes.html#str
[699] https://docs.python.org/3/library/stdtypes.html#dict
[700] https://docs.python.org/3/library/stdtypes.html#str

- **name_classes** (*tuple*[701]) – name classes to save, only the scientific and common name are stored

Returns dictionary with merged_id -> taxon_id

Return type dict[702]

mgkit.taxon.**parse_ncbi_taxonomy_nodes_file**(*file_handle*, *taxa_names=None*)
New in version 0.2.3.

Parses the *nodes.dmp* file where the nodes of the taxonomy are stored. Available at ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/.

Parameters

- **file_handle** (*str*[703], *file*) – file name or handle to the file
- **taxa_names** (*dict*[704]) – dictionary with the taxa names (returned from *parse_ncbi_taxonomy_names_file()*)

Yields *UniprotTaxonTuple* – UniprotTaxonTuple instance

mgkit.taxon.**parse_uniprot_taxon**(*line*, *light=True*)
Changed in version 0.1.13: now accepts empty scientific names, for root taxa

Changed in version 0.2.1: added *light* parameter

Parses a Uniprot taxonomy file (tab delimited) line and returns a UniprotTaxonTuple instance. If *light* is True, lineage is not stored to decrease the memory usage. This is now the default.

mgkit.taxon.**taxa_distance_matrix**(*taxonomy*, *taxon_ids*)
New in version 0.1.16.

Given a list of taxonomic identifiers, returns a distance matrix in a pairwise manner by using *distance_two_taxa()* on all possible two element combinations of *taxon_ids*.

Parameters

- **taxonomy** – *UniprotTaxonomy* instance
- **taxon_ids** (*iterable*) – list taxonomic identifiers

Returns matrix with the pairwise distances of all *taxon_ids*

Return type pandas.DataFrame

7.1.3 Module contents

Metagenomics Framework

exception mgkit.**DependencyError**(*package*)
Bases: exceptions.Exception

Raised if an optional requirement is needed

mgkit.**check_version**(*version*)

mgkit.**cite**(*file_handle=<open file '<stderr>', mode 'w'>*)
Print citation to the specified stream

7.2 mgkit

[701] https://docs.python.org/3/library/stdtypes.html#tuple
[702] https://docs.python.org/3/library/stdtypes.html#dict
[703] https://docs.python.org/3/library/stdtypes.html#str
[704] https://docs.python.org/3/library/stdtypes.html#dict

CHAPTER 8

Changes

8.1 0.3.1

This release adds several scripts and commands. Successive releases 0.3.x releases will be used to fix bugs and refine the APIs and CLI. Most importantly, since the publishing of the first paper using the framework, the releases will go torward the removal of as much deprecated code as possible. At the same time, a general review of the code to be able to run on Python3 (probably via the *six* package) will start. The general idea is to reach as a full removal of legacy code in 0.4.0, while full Python3 compatibility is the aim of 0.5.0, which also means dropping dependencies that are not compatible with Python3.

8.1.1 Added

- *mgkit.graphs.from_kgml()* to make a graph from a KGML file (allows for directionality)
- *mgkit.graphs.add_module_compounds()*: updates a graph with compounds information as needed
- *mgkit.kegg.parse_reaction()*: parses a reaction equation from Kegg
- added *–no-frame* option to *hmmer2gff - Convert HMMER output to GFF*, to use non translated protein sequences. Also changed the *mgkit.io.gff.from_hmmer()* function to enable this behaviour
- added options *–num-gt* and *–num-lt* to the *values* command of *filter-gff - Filter GFF annotations* to filter based on > and < inequality, in addition to >= and <=
- added *uid* as command in *fasta-utils - Fasta Utilities* to make unique fasta headers
- methods to make `mgkit.db.mongo.GFFDB` to behave like a dictionary (an annotation *uid* can be used as a key to retrieve it, instead of a query), this includes the possibility to iterate over it, but what is yielded are the values, not the keys (i.e. *mgkit.io.gff.Annotation* instances, not *uid*)
- added *mgkit.counts.func.from_gff()* to load count data stored inside a GFF, as is the case when the *counts* command of *add-gff-info - Add informations to GFF annotations* is used'
- added *mgkit.kegg.KeggClientRest.conv()* and *mgkit.kegg.KeggClientRest.find()* operations to *mgkit.kegg.KeggClientRest*
- *mgkit.kegg.KeggClientRest* now caches calls to several methods. The cache can be written to disk using *mgkit.kegg.KeggClientRest.write_cache()* or emptied via *mgkit.kegg.KeggClientRest.empty_cache()*

- added *mgkit.utils.dictionary.merge_dictionaries()* to merge multiple dictionaries where the keys contain different values
- added a Docker file to make a preconfigured mgkit/jupyter build
- added C functions (using Cython) for tetramer/kmer counting. The C functions are the default, with the pure python implementation having a _ appended to their names. This is because the Cython functions cannot have docstrings
- added *mgkit.io.gff.annotation_coverage_sorted()*
- added *mgkit.io.gff.Annotation.to_dict()*
- added *mgkit.plots.utils.legend_patches()* to create matplotlib patches, to be in legends
- added scripts download IDs to taxa tables from NCBI/Uniprot
- added *mgkit.io.utils.group_tuples_by_key()*
- added *cov* command to *get-gff-info - Extract informations to GFF annotations* and *filter-gff - Filter GFF annotations*
- added *mgkit.io.fasta.load_fasta_prodigal()*, to load the fasta file from prodigal for called genes (tested on aminoacids)
- added option to output a JSON file to the *lca* command in ref:*taxon-utils* and *cov* command in *get-gff-info - Extract informations to GFF annotations*
- added a bash script, *sort-gff.sh* to help sort a GFF
- added *mgkit.taxon.UniprotTaxonomy.get_lineage()* which simplifies the use of *mgkit.taxon.get_lineage()*
- added *mgkit.io.fastq.load_fastq()* as a simple parser for fastq files
- added a new script, *sampling-utils - Resampling Utilities*
- added *mgkit.utils.common.union_ranges()* and *mgkit.utils.common.complement_ranges()*
- added *to_hdf* command to *taxon-utils - Taxonomy Utilities* to create a HDF5 file to lookup taxa tables from NCBI/Uniprot
- added *–hdf-table* option to *addtaxa* command in *add-gff-info - Add informations to GFF annotations*
- *mgkit.taxon.UniprotTaxonomy.add_taxon()*, *mgkit.taxon.UniprotTaxonomy.add_lineage()* and *mgkit.taxon.UniprotTaxonomy.drop_taxon()*

8.1.2 Changed

- changed *domain* to *superkingdom* as for NCBI taxonomy in *mgkit.taxon.UniprotTaxonomy.read_from_gtdb_taxonomy()*
- updated scripts documentation to include installed but non advertised scripts (like *translate_seq - Translate Nucleotides to Aminoacids (Deprecated)*)
- *mgkit.kegg.KeggReaction* was reworked to only store the equation information
- some commands in *fastq-utils - Fastq Utilities* did not support standard in/out, also added the script usage to the script details
- *translate_seq - Translate Nucleotides to Aminoacids (Deprecated)* now supports standard in/out
- added *haplotypes* parameter to *mgkit.snps.funcs.combine_sample_snps()*
- an annotation from `mgkit.db.mongo.GFFDB` now doesn't include the lineage, because it conflicts with the string used in a GFF file
- an *mgkit.io.gff.Annotation.coverage()* now returns a *float* instead od a *int*

- moved code from package *mgkit.io* to *mgkit.io.utils*
- changed behaviour of *mgkit.utils.common.union_range()*
- removed mgkit.utills.common.range_substract_()
- added *progressbar2* as installation requirement
- changed how *mgkit.taxon.UniprotTaxonomy.find_by_name()*

8.1.3 Fixed

Besides smaller fixes:

- *mgkit.plots.abund.draw_circles()* behaviour when *sizescale* doesn't have the same shape as *order*
- parser is now correct for *taxon-utils - Taxonomy Utilities*, to include the Krona[705] options
- ondition when a blast output is empty, hence *lineno* is not initialised when a message is logged

8.1.4 Deprecated

- *translate_seq - Translate Nucleotides to Aminoacids (Deprecated)* will be removed in version 0.4.0, instead use the similar command in *fasta-utils - Fasta Utilities*

8.2 0.3.0

A lot of bugs were fixed in this release, especially for reading NCBI taxonomy and using the *msgpack* format to save a UniprotTaxonomy instance. Also added a tutorial for profiling a microbial community using MGKit and BLAST (*Profile a Community with BLAST*)

8.2.1 Added

- *mgkit.align.read_samtools_depth()* to read the samtools depth format iteratively (returns a generator)
- *mgkit.align.SamtoolsDepth*, used to cache the samtools depth format, while requesting region coverage
- *mgkit.kegg.KeggModule.find_submodules()*, *mgkit.kegg.KeggModule.parse_entry2()*
- *mgkit.mappings.enzyme.get_mapping_level()*
- *mgkit.utils.dictionary.cache_dict_file()* to cache a large dictionary file (tab separated file with 2 columns), an example of its usage is in the documentation
- *mgkit.taxon.UniprotTaxonomy.read_from_gtdb_taxonomy()* to read a custom taxonomy from a tab separated file. The taxon_id are not guaranteed to be stable between runs
- added *cov_samtools* to *add-gff-info* script
- added *mgkit.workflow.fasta_utils* and correspondent script *fasta-utils*
- added options *-k* and *-kt* to *taxon_utils*, which outputs a file that can be used with Krona *ktImportText* (needs to use *-q* with this script)

[705] https://github.com/marbl/Krona/wiki

8.2.2 Changed

- added *no_zero* parameter to `mgkit.io.blast.parse_accession_taxa_table()`
- changed behaviour of `mgkit.kegg.KeggModule` and some of its methods.
- added *with_last* parameter to `mgkit.taxon.get_lineage()`
- added –*split* option to *add-gff-info exp_syn* and *get-gff-info sequence* scripts, to emulate BLAST behaviour in parsing sequence headers
- added *-c* option to *add-gff-info addtaxa*

8.3 0.2.5

8.3.1 Changed

- added the *only_ranked* argument to `mgkit.taxon.get_lineage()`
- *add-gff-info addtaxa* (*add-gff-info - Add informations to GFF annotations*) doesn't preload the GFF file if a dictionary is used instead of the taxa table
- *blast2gff blastdb* ((*blast2gff - Convert BLAST output to GFF*) offers more options to control the format of the header in the DB used
- added the *sequence* command to *filter-gff* (*filter-gff - Filter GFF annotations*), to filter all annotations on a per-sequence base, based on mean bitscore or other comparisons

8.3.2 Added

- added `mgkit.counts.func.load_counts_from_gff()`
- added `mgkit.io.blast.parse_accession_taxa_table()`
- added `mgkit.plots.abund.draw_axis_internal_triangle()`
- added representation of `mgkit.taxon.UniprotTaxonomy`, it show the number of taxa in the instance
- added `mgkit.taxon.last_common_ancestor_multiple()`
- added *taxon_utils* (*taxon-utils - Taxonomy Utilities*) to filter GFF based on their taxonomy and find the last common ancestor for a reference sequence based on either GFF annotations or a list of taxon_ids (in a text file)

8.4 0.2.4

8.4.1 Changed

- `mgkit.utils.sequence.get_contigs_info()` now accepts a dictionary name->seq or a list of sequences, besides a file name (r536)
- *add-gff-info* **counts** command now removes trailing commas from the samples list
- the axes are turned off after the dendrogram is plo

8.4.2 Fixed

- the **snp_parser** script requirements were set wrong in *setup.py* (r540)
- uncommented lines to download sample data to build documentation (r533)
- *add-gff-info* **uniprot** command now writes the *lineage* attribute correctly (r538)

8.5 0.2.3

The installation dependencies are more flexible, with only *numpy* as being **required**. To install every needed packages, you can use:

```
$ pip install mgkit[full]
```

8.5.1 Added

- new option to pass the *query sequences* to **blast2gff**, this allows to add the correct frame of the annotation in the GFF
- added the attributes *evalue*, *subject_start* and *subject_end* to the output of *blast2gff*. The subject start and end position allow to understand on which frame of the *subject sequence* the match was found
- added the options to annotate the heatmap with the numbers. Also updated the relative example notebook
- Added the option to reads the taxonomy from NCBI dump files, using `mgkit.taxon.UniprotTaxonomy.read_from_ncbi_dump()`. This make it faster to get the taxonomy file
- added argument to return information from `mgkit.net.embl.datawarehouse_search()`, in the form of tab separated data. The argument *fields* can be used when *display* is set to **report**. An example on how to use it is in the function documentation
- added a bash script *download-taxonomy.sh* that download the taxonomy
- added script *venv-docs.sh* to build the documentation in HTML under a virtual environment. matplotlib on MacOS X raises a RuntimeError, because of a bug in virtualenv[706], the documentation can be first build with this, after the script *create-apidoc.sh* is create the API documentation. The rest of the documentation (e.g. the PDF) can be created with *make* as usual, afterwards
- added `mgkit.net.pfam`, with only one function at the moment, that returns the descriptions of the families.
- added *pfam* command to *add-gff-info*, using the mentioned function, it adds the description of the Pfam families in the GFF file
- added a new exception, used internally when an additional dependency is needed

8.5.2 Changed

- using the NCBI taxonomy dump has two side effects:
  - the scientific/common names are kept as is, not lower cased as was before
  - a *merged* file is provided for *taxon_id* that changed. While the old taxon_id is kept in the taxonomy, this point to the new taxon, to keep backward compatibility
- renamed the *add-gff-info gitaxa* command to *addtaxa*. It now accepts more data sources (dictionaries) and is more general

[706] https://github.com/pypa/virtualenv/issues/54

- changed `mgkit.net.embl.datawarehouse_search()` to automatically set the limit at 100,000 records
- the taxonomy can now be saved using msgpack[707], making it faster to read/write it. It's also more compact and better compression ratio
- the `mgkit.plots.heatmap.grouped_spine()` now accept the rotation of the labels as option
- added option to use another attribute for the *gene_id* in the *get-gff-info* script *gtf* command
- added a function to compare the version of MGKit used, throwing a warning, when it's different (`mgkit.check_version()`)
- removed test for old SNPs structures and added the same tests for the new one
- `mgkit.snps.classes.GeneSNP` now caches the number of synonymous and non-synonymous SNPs for better speed
- `mgkit.io.gff.GenomicRange.__contains__()` now also accepts a tuple (start, end) or another GenomicRange instance

8.5.3 Fixed

- a bug in the *gitaxa* (now *addtaxa*) command: when a taxon_id was not found in the table, the wrong *taxon_name* and *lineage* was inserted
- bug in `mgkit.snps.classes.GeneSNP` that prevented the correct addition of values
- fixed bug in `mgkit.snps.funcs.flat_sample_snps()` with the new class
- `mgkit.io.gff.parse_gff()` now correctly handles comment lines and stops parsing if the fasta file at the end of a GFF is found

8.6 0.2.2

8.6.1 Added

- new commands for the **add-gff-info** script (*add-gff-info - Add informations to GFF annotations*):
  - *eggnog* to add information from eggNOG HMMs (at the moment the 4.5 Viral)
  - *counts* and *fpkms* to add count data (correctly exported to mongodb)
  - *gitaxa* to add taxonomy information directly from GI identifiers from NCBI
- added *blastdb* command to **blast2gff** script (*blast2gff - Convert BLAST output to GFF*)
- updated *MGKit GFF Specifications*
- added *gtf* command to **get-gff-info** script (*get-gff-info - Extract informations to GFF annotations*) to convert a GFF to GTF, that is accepted by featureCounts[708], in conjunction with the *counts* command of **add-gff-info**
- added method to `mgkit.snps.classes.RatioMixIn.calc_ratio_flag` to calculate special cases of pN/pS

[707] https://github.com/msgpack/msgpack-python
[708] http://bioinf.wehi.edu.au/featureCounts/

8.6.2 Changed

- added argument in functions of the *mgkit.snps.conv_func* to bypass the default filters
- added *use_uid* argument to *mgkit.snps.funcs.combine_sample_snps()* to use the *uid* instead of the *gene_id* when calculating pN/pS
- added *flag_values* argument to *mgkit.snps.funcs.combine_sample_snps()* to use *mgkit.snps.classes.RatioMixIn.calc_ratio_flag* instead of *mgkit.snps.classes.RatioMixIn.calc_ratio*

8.6.3 Removed

- deprecated code from the **snps** package

8.7 0.2.1

8.7.1 Added

- added `mgkit.db.mongo`
- added `mgkit.db.dbm`
- added *mgkit.io.gff.Annotation.get_mappings()*
- added *mgkit.io.gff.Annotation.to_json()*
- added *mgkit.io.gff.Annotation.to_mongodb()*
- added *mgkit.io.gff.from_json()*
- added *mgkit.io.gff.from_mongodb()*
- added *mgkit.taxon.get_lineage()*
- added *mgkit.utils.sequence.get_contigs_info()*
- added *mongodb* and *dbm* commands to script *get-gff-info*
- added *kegg* command to *add-gff-info* script, caching results and *-d* option to *uniprot* command
- added *-ft* option to *blast2gff* script
- added *-ko* option to *download_profiles*
- added new HMMER tutorial
- added another notebook to the plot examples, for misc. tips
- added a script that downloads from figshare the tutorial data]
- added function to get an enzyme full name (*mgkit.mappings.enzyme.get_enzyme_full_name()*)
- added example notebook for using GFF annotations and the `mgkit.db.dbm`, `mgkit.db.mongo` modules

8.7.2 Changed

- *mgkit.io.blast.parse_uniprot_blast()*
- *mgkit.io.gff.Annotation*
- *mgkit.io.gff.GenomicRange*

- *mgkit.io.gff.from_hmmer()*
- *mgkit.taxon.UniprotTaxonomy.read_taxonomy()*
- *mgkit.taxon.parse_uniprot_taxon()*
- changed behaviour of *hmmer2gff* script
- changed tutorial notebook to specify the directory where the data is

8.7.3 Deprecated

- mgkit.filter.taxon.filter_taxonomy_by_lineage()
- mgkit.filter.taxon.filter_taxonomy_by_rank()

8.7.4 Removed

- removed old *filter_gff* script

8.8 0.2.0

- added creation of wheel distribution
- changes to ensure compatibility with alter pandas versions
- *mgkit.io.gff.Annotation.get_ec()* now returns a set, reflected changes in tests
- added a *–cite* option to scripts
- fixes to tutorial
- updated documentation for sphinx 1.3
- changes to diagrams
- added decoration to raise warnings for deprecated functions
- added possibility for *mgkit.counts.func.load_sample_counts()* info_dict to be a function instead of a dictionary
- consolidation of some eggNOG structures
- added more spine options in *mgkit.plots.heatmap.grouped_spine()*
- added a *length* property to *mgkit.io.gff.Annotation*
- changed *filter-gff* script to customise the filtering function, from the default one, also updated the relative documentation
- fixed a few plot functions

8.9 0.1.16

- changed default parameter for *mgkit.plots.boxplot.add_values_to_boxplot()*
- Added *include_only* filter option to the default snp filters mgkit.consts.DEFAULT_SNP_FILTER
- the default filter for SNPs now use an include only option, by default including only protozoa, archaea, fungi and bacteria in the matrix
- added *widths* parameter to def *mgkit.plots.boxplot.boxplot_dataframe()* function, added function *mgkit.plots.boxplot.add_significance_to_boxplot()* and updated example boxplot notebook for new function example

- *use_dist* and *dist_func* parameters to the `mgkit.plots.heatmap.dendrogram()` function
- added a few constants and functions to calculate the distance matrices of taxa: `mgkit.taxon.taxa_distance_matrix()`, `mgkit.taxon.distance_taxa_ancestor()` and `mgkit.taxon.distance_two_taxa()`
- `mgkit.kegg.KeggClientRest.link_ids()` now accept a dictionary as list of ids
- if the conversion of an Annotation attribute (first 8 columns) raises a ValueError in `mgkit.io.gff.from_gff()`, by default the parser keeps the string version (cases for phase, where is '.' instead of a number)
- treat cases where an attribute is set with no value in `mgkit.io.gff.from_gff()`
- added `mgkit.plots.colors.palette_float_to_hex()` to convert floating value palettes to string
- forces vertical alignment of tick labels in heatmaps
- added parameter to get a consensus sequence for an AA alignment, by adding the *nucl* parameter to `mgkit.utils.sequence.Alignment.get_consensus()`
- added `mgkit.utils.sequence.get_variant_sequence()` to get variants of a sequence, essentially changing the sequence according to the SNPs passed
- added method to get an aminoacid sequence from Annotation in `mgkit.io.gff.Annotation.get_aa_seq()` and added the possibility to pass a SNP to get the variant sequence of an Annotation in `mgkit.io.gff.Annotation.get_nuc_seq()`.
- added *exp_syn* command to *add-gff-info* script
- changed GTF file conversion
- changed behaviour of `mgkit.taxon.is_ancestor()`: if a *taxon_id* raises a KeyError, False is now returned. In other words, if the taxon_id is not found in the taxonomy, it's not an ancestor
- added `mgkit.io.gff.GenomicRange.__contains__()`. It tests if a position is inside the range
- added `mgkit.io.gff.GenomicRange.get_relative_pos()`. It returns a position relative to the GenomicRange start
- fixed documentation and bugs (Annotation.get_nuc_seq)
- added `mgkit.io.gff.Annotation.is_syn()`. It returns True if a SNP is synonymous and False if non-synonymous
- added *to_nuc* parameter to `mgkit.io.gff.from_nuc_blast()` function. It to_nuc is False, it is assumed that the hit was against an amino acidic DB, in which case the phase should always set to 0
- reworked internal of *snp_parser* script. It doesn't use SNPDat anymore
- updated tutorial
- added ipython notebook as an example to explore data from the tutorial
- cleaned deprecated code, fixed imports, added tests and documentation

8.10 0.1.15

- changed name of `mgkit.taxon.lowest_common_ancestor()` to `mgkit.taxon.last_common_ancestor()`, the old function name points to the new one
- added `mgkit.counts.func.map_counts_to_category()` to remap counts from one ID to another
- added *get-gff-info* script to extract information from GFF files
- script *download_data* can now download only taxonomy data

- added more script documentation
- added examples on gene prediction
- added function *mgkit.io.gff.from_hmmer()* to parse HMMER results and return *mgkit.io.gff.Annotation* instances
- added *mgkit.io.gff.Annotation.to_gtf()* to return a GTF line, *mgkit.io.gff.Annotation.add_gc_content()* and *mgkit.io.gff.Annotation.add_gc_ratio()* to calculate GC content and ratio respectively
- added *mgkit.io.gff.parse_gff_files()* to parse multiple GFF files
- added *uid_used* parameter to several functions in *mgkit.counts.func*
- added *mgkit.plots.abund* to plot abundance plots
- added example notebooks for plots
- HTSeq is now required only by the scripts that uses it, *snp_parser* and *fastq_utils*
- added function to convert numbers when reading from htseq count files
- changed behavior of *-b* option in *add-gff-info taxonomy* command
- added *mgkit.io.gff.get_annotation_map()*

8.11 0.1.14

- added ipthon notebooks to the documentation. As of this version the included ones (in *docs/source/examples*) are for two plot modules. Also added a bash script to convert them into rst files to be included with the documentation. The *.rst* are not versioned, and they must be rebuild, meaning that one of the requirements for building the docs is to have IPython[709] installed with the notebook extension
- now importing some packages automatically import the subpackages as well
- refactored *mgkit.plots* into a package, with most of the original functions imported into it, for backward compatibility
- added *mgkit.graphs.build_weighted_graph()*
- added *box_vert* parameter in *mgkit.plots.boxplot.add_values_to_boxplot()*, the default will be changed in a later version (kept for compatibility with older scripts/notebooks)
- added an heatmap module to the plots package. Examples are in the notebook
- added *mgkit.align.covered_annotation_bp()* to find the number of bp covered by reads in annotations (as opposed to using the annotation length)
- added documentation to *mgkit.mappings.eggnog.NOGInfo* and an additional method
- added *mgkit.net.uniprot.get_uniprot_ec_mappings()* as it was used in a few scripts already
- added *mgkit.mappings.enzyme.change_mapping_level()* and other to deal with EC numbers. Also improved documentation with some examples
- added *mgkit.counts.func.load_sample_counts_to_genes()* and *mgkit.counts.func.load_sample_counts_to_taxon()*, for mapping counts to only genes or taxa. Also added *index* parameter in *mgkit.counts.func.map_counts()* to accomodate the changes
- added *mgkit.net.uniprot.get_ko_to_eggnog_mappings()* to get mappings of KO identifiers to eggNOG
- added *mgkit.io.gff.split_gff_file()* to split a gff into several ones, assuring that all annotations for a sequence is in the same file; useful to split massive GFF files before filtering

[709] http://ipython.org

- added *mgkit.counts.func.load_deseq2_results()* to load DESeq2 results in *CSV* format
- added *mgkit.counts.scaling.scale_rpkm()* for scale with rpkm a count table
- added caching options to *mgkit.counts.func.load_sample_counts()* and others
- fixes and improvements to documentation

8.12 0.1.13

- added counts package, including functions to load HTSeq-counts results and scaling
- added *mgkit.filter.taxon.filter_by_ancestor()*, as a convenience function
- deprecated functions in *mgkit.io.blast* module, added more to parse blast outputs (some specific)
- *mgkit.io.fasta.load_fasta()* returns uppercase sequences, added a function (*mgkit.io.fasta.split_fasta_file()*) to split fasta files
- added more methods to *mgkit.io.gff.Annotation* to complete API from old annotations
- fixed *mgkit.io.gff.Annotation.dbq* property to return an **int** (bug in filtering with filter-gff)
- added function to extract the sequences covered by annotations, using the *mgkit.io.gff.Annotation.get_nuc_seq()* method
- added `mgkit.io.gff.correct_old_annotations()` to update old annotated GFF to new conventions
- added *mgkit.io.gff.group_annotations_by_ancestor()* and *mgkit.io.gff.group_annotations_sorted()*
- moved deprecated GFF classes/modules in `mgkit.io.gff_old`
- added *mgkit.io.uniprot* module to read/write Uniprot files
- added *mgkit.kegg.KeggClientRest.get_ids_names()* to remove old methods to get specific class names used to retrieve (they are deprecated at the moment)
- added *mgkit.kegg.KeggModule* to parse a Kegg module entry
- added *mgkit.net.embl.datawarehouse_search()* to search EMBL resources
- made *mgkit.net.uniprot.query_uniprot()* more flexible
- added/changed plot function in *mgkit.plots*
- added enum34 as a dependency for Python versions below 3.4
- changed classes to hold SNPs data: deprecated `mgkit.snps.classes.GeneSyn`, replaced by *mgkit.snps.classes.GeneSNP* which the enum module for *mgkit.snps.classes.SNPType*
- added *mgkit.taxon.NoLcaFound*
- fixed behaviour of *mgkit.taxon.UniprotTaxonomy.get_ranked_taxon()* for newer taxonomies
- change behaviour of *mgkit.taxon.UniprotTaxonomy.is_ancestor()* to use module *mgkit.taxon.is_ancestor()* and accept multiple taxon IDs to test
- *mgkit.taxon.UniprotTaxonomy.load_data()* now accept compressed data and file handles
- added *mgkit.taxon.lowest_common_ancestor()* to find the lowest common ancestor of two taxon IDs
- changed behaviour of *mgkit.taxon.parse_uniprot_taxon()*
- added functions to get GC content, ratio of a sequence and it composition to *mgkit.utils.sequence*
- added more options to **blast2gff** script

- added *coverage*, *taxonomy* and *unipfile* to **add-gff-info**
- refactored **snp_parser** to use new classes
- added possibility to use sorted GFF files as input for **filter-gff** to use less memory (the examples show how to use *sort* in Unix)

8.13 0.1.12

- added functions to elongate annotations, measure the coverage of them and diff GFF files in *mgkit.io.gff*
- added ranges_length and union_ranges to *mgkit.utils.common*
- added script filter-gff, filter_gff will be deprecated
- added script blast2gff to convert blast output to a GFF
- removed unneeded dependencies to build docs
- added script add-gff-info to add more annotations to GFF files
- added *mgkit.io.blast.parse_blast_tab()* to parse BLAST tabular format
- added *mgkit.io.blast.parse_uniprot_blast()* to return annotations from a BLAST tabular file
- added `mgkit.graph` module
- added classes *mgkit.io.gff.Annotation* and *mgkit.io.gff.GenomicRange* and deprecated old classes to handle GFF annotations (API not stable)
- added *mgkit.io.gff.DuplicateKeyError* raised in parsing GFF files
- added functions used to return annotations from several sources
- added option *gff_type* in `mgkit.io.gff.load_gff()`
- added *mgkit.net.embl.dbfetch()*
- added *mgkit.net.uniprot.get_gene_info()* and *mgkit.net.uniprot.query_uniprot()* *mgkit.net.uniprot.parse_uniprot_response()*
- added apply_func_to_values to *mgkit.utils.dictionary*
- added *mgkit.snps.conv_func.get_full_dataframe()*, *mgkit.snps.conv_func.get_gene_taxon_dataframe()*
- added more tests

8.14 0.1.11

- removed *rst2pdf* for generating a PDF for documentation. Latex is preferred
- corrections to documentation and example script
- removed need for joblib library in *translate_seq* script: used only if available (for using multiple processors)
- deprecated `mgkit.snps.funcs.combine_snps_in_dataframe()` and `mgkit.snps.funcs.combine_snps_in_dataframe()`: *mgkit.snps.funcs.combine_sample_snps()* should be used
- refactored some tests and added more
- added *docs_req.txt* to help build the documentation ont readthedocs.org

- renamed `mgkit.snps.classes.GeneSyn` gid and taxon attributes to gene_id and taxon_id. The old names are still available for use (via properties), but the will be taken out in later versions. Old pickle data should be loaded and saved again before in this release
- added a few convenience functions to ease the use of *combine_sample_snps()*
- added function *mgkit.snps.funcs.significance_test()* to test the distributions of genes share between two taxa.
- fixed an issue with deinterleaving sequence data from khmer
- added *mgkit.snps.funcs.flat_sample_snps()*
- Added method to *mgkit.kegg.KeggClientRest* to get names for all ids of a certain type (more generic than the various *get_*_names*)
- added first implementation of *mgkit.kegg.KeggModule* class to parse a Kegg module entry
- *mgkit.snps.conv_func.get_rank_dataframe()*, *mgkit.snps.conv_func.get_gene_map_dataframe()*

CHAPTER 9

Indices and tables

- genindex
- modindex
- search

Python Module Index

m

Index

Symbols

A

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W