BioBundle v1.1- User Manual

Carsten Kemena

December 10, 2017

Contents

1 seqExtract

`seqExtract` is a small program to extract sequences from a sequence set. It allows to extract sequences many different criteria.

1.1 Options

General options:

Table 1.1: General options for *seqExtract*

option	description
-h –help	Displays a simple help message
-i –in	The input file. Currently only the `fasta`-format is supported.
-I –index	Uses an index file. This file can speed up sequence extraction in larger files (e.g. genomes) especially when rerunning the command several times. One can specify the path to the index file using the 'indexFile' parameter. If none is specified a file will be created with same name as input file but with an additional '.sei' extension.
-F –indexFile	The path to the index file if not the default position should be used.
-l –inputList	A file containing a list of files to be used as input

Output options:

Table 1.2: Output options for *seqExtract*

Option	Description
-o, –out	The output file. Currently only `fasta`-format output is supported.
-a, –append	Appends the extracted sequences to a file. (default: false)
-c, –remove-comments	Remove comments from the fasta headers. (default: false)

Extract options:

Table 1.3: Extract options for *seqExtract*

Option	default	Description
-e, –extract	-	The sequence to extract
-d, –delimiter	-	The delimiter to use
-E, –regex	-	A regular expression to match a name
-r, –remove	-	Remove the given sequences
-n, –numSeqs	-	The number of sequences to extract
-s, –seed	-	Seed for random extract function
-f, –function	-	The function to use for extraction
-m, –ignore-missing	-	Ignore missing sequences

Table 1.4: Modifying options for *seqExtract*

Option	Default value	Description
-t, –translate	-	Translate into amino acid
-T, –table	arg (=standard) -	The translation table to use
-R, –revComp	-	Calculate the reverse complement

2 isoformCleaner

`isoformCleaner` is a small program to remove isoforms from a set and keep only the largest one. This is often useful as for many analyses isoforms will influence the results.

2.1 Options

This section explains all the parameters that can be used with the program.

2.1.1 General options

These are the general options that can be used.

Table 2.1: General options for *seqExtract*

option	default	description
option	default	effect
-h, –help	-	Prints a simple help message
-i, –in	-	The sequence file to clean
-o, –out	-	The output file
-s, –splitchar	-	The split character to use to distinguish gene name from isoform extension (e.g. - in Gene1-PA).

2.1.2 Regex options

These are the options that can be used for a regular expression based cleaning. The regular expression should identify a section of the sequence header that is the same for all isoforms of a gene (e.g. the gene name).

Table 2.2: General options for *seqExtract*

option	default	description
option	default	effect
-r, –regular	-	Regular expression
-c, –comment	-	Search comment only
-n, –name	-	Search name only
-p, –preset	-	Preset regex.

For two common patterns some regular expressions have precoded. The availabele presets can be seend in the table below.

Table 2.3: Preset options

name	regular expression
gene	`gene[:=]\s*([\S]+)`
flybase	`parent=(FBgn[^ ,]+,)`

2.1.3 GFF options

If none of the above works there is the chance that you can use a gff to do the isoform cleaning. This is highly experimental and only works if the Parent field is present. Furthermore currently there is no support for multiple parents.

Table 2.4: GFF options

option	default	description
-g, –gff	-	The gff file in gff3 format.
-t, –type	mRNA	The feature type that contains the sequence name as ID and the gene name in the parent field.
-f, –id-field	ID	The argument is used to identify the field in the GFF-file that contains the sequence names. Usually ID is correct.

2.2 Examples

Here are some very basic example on the usage of isoform cleaner.

```
# use a split character
isoformCleaner -s '.' -i foo.fa -o bar.fa

# usa a regular expression
isoformCleaner -r "parent=(FBgn[^ ,]+,)" -i foo.fa -o bar.fa

# same as above but with a preset value
isoformCleaner -p flybase -i foo.fa -o bar.fa

# gff cleaning
isoformCleaner -i foo.fa -g bar.gff
```

3 stopCleaner

`stopCleaner` can remove stop characters at the end of a sequence and furthermore allows to remove pseudogenes (as defined by having a stop codon in the middle of the sequence) from a set.

Table 3.1: General options for *seqExtract*

option	default	description
-h, –help	-	Produces a simple help message
-i, –in	-	The input file: Protein sequences in fasta format. If none is provided sequences are read from stdin.
-o, –out	-	The output file. If none is provided sequences will be printed to stdout
–no-final-stop-removal	-	Do not remove the final stop characters.
-r, –remove-pseudogenes	-	Remove pseudogenes. Pseudogenes are in this case all genes that do not contain a stop at the end.
-s, –stop	.*	The stop characters to use.
-l, –list	-	Produces a list of genes that were removed and writes it to the provided file.
-k, –keep	-	Keep sequences with these IDs. Can be useful to prevent removal of sequences that are known to correctly contain a stop codon (etc. selenoproteins)
-R, –replace	-	Replace stops with this chararcter.

Examples:

```
./stopCleaner -i dmel.fa
```