BioBundle v1.1- User Manual

Carsten Kemena

March 14, 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 value	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.

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 pr vided sequences are expected to be provided via a pipe.
-o, –out	-	The output file. If none is provided sequences will be printed t stdout
–no-final-stop-removal	-	Do not remove the final stop characters.
-r, –remove-pseudogenes	-	Remove pseudogenes. Pseudogenes are in this case all genes tha 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 th provided file.
-k, –keep	-	Keep sequences with these IDs. Can be useful to prevent remov of sequences that are known to correctly contain a stop codon (et selenoproteins)
-R, –replace	-	Replace stops with this chrararcter.

Examples:

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