Eslam S. Ibrahim

SDDC Cheat sheet – 2018
https://sites.google.com/pharma.cu.edu.eg/eslam-ibrahim

eslam.ebrahim@pharma.cu.edu.eg
eslam_samir_ragab.ibrahim@stud-mail.uni-wuerzburg.de

SDDC Cheat sheet

- SDDC works in all platforms and with different python versions. To overcome all of these varieties, “future” package was added to the source code.

> *Note*: you may need to update future module on your computer by following those commands:
>
> 1. Download get-pip (https://bootstrap.pypa.io/get-pip.py)
> 2. `python get-pip.py install future`
> 3. `python get-pip.py install future --upgrade`

- The default working directory is the directory where your sddc.py located.
- You can use any file in its own directory by specifying the directory before the file’s name

```
-in G:\eslam\input.txt –out D:\samir\out.txt
```

Dereplication Mode:

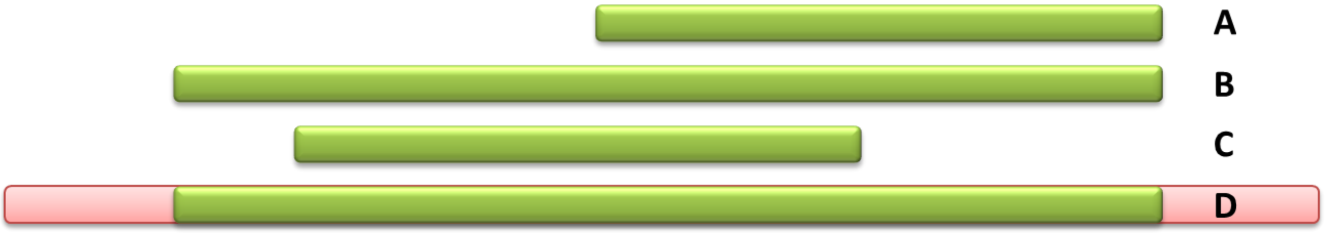
- Dereplication mode removes the 100% exact sequences and the shorter partial sequences.
- You can retain the original order of the sequences by adding the following option `-org_order` to you command line.
- You can use it for one file or multiple files (with the same extension for which the program will ask you and in the same directory).
- You can use it on fasta file or fastq file.

> *Note*: if you dereplicate a fastq file, you will not be able to perform some statistics on the resulted file.

- You will have an output file called (**names_of_deleted.txt**) that contains the dereplicated names of deleted sequences.

> *Note*: if the **names_of_deleted.txt** contains less than the numbers of deleted sequences, you have some exactly replicated names.

- Optimum length approach needs an approximate of your ideal protein length (the program slides ±10% around your ideal length).



Largest possible length: the program keeps the “D” sequence while deleting “A”, “B” and “C”.

Optimum length: the program keeps the “B” sequence while deleting “A”, “C” and “D”.

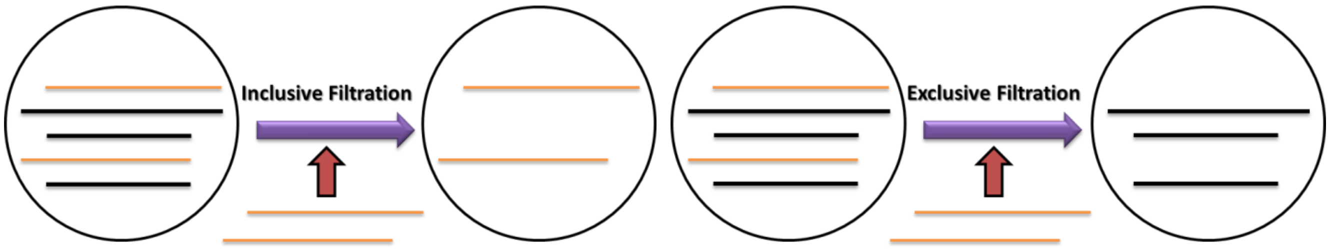
- You can set a minimum length option if you want to have a minimum length of your sequences.

Note (1): Derep mode in Nucleotide sequences takes longer time due to additional searching for reverse complement replicates.

Note (2): you can use the **names_of_deleted.txt** to reuse SDDC in filter mode to filter your original database inclusively to obtain all the replicated sequences for more inspection.

Filtration Mode:

- For name filtration: the name should be FASTA – formatted even if there are no sequences in the file. i.e. begins with (>) and ends with new paragraph character.
- **Update:** you can filter now using keywords only without the whole FASTA header by adding `-kw` argument to your command line.
- The output of filtration mode always contains non-redundant sequences (exact match) but will retain partial sequences.
- Inclusive or Exclusive approach could only be applied when filter by name, while sequence filtration is always exclusive.



Hint: you can use the same filtration file in both (by name and by seq) if it is regularly fasta formatted.

Exchange headers Mode:

- Exchange headers mode changes the headers of your FASTA file with any provided alternative
- The exchange file should be in csv format with the old words in (column 1) and the new words in (column 2).
- For example, if your exchange file like the following:
  aaa,bbb
  ccc,ddd
  then each occurrence of (aaa) in any FASTA header in your input file will be replaced with (bbb) as a replacement.

Input file	Output file
>aaa	>bbb
ATCGATCG	ATCGATCG
>cccaaa	>dddbbb
TCGATCGA	TCGATCGA

- Your csv file shouldn’t have column headers.
- The old and new words are case-sensitive (upper and lower cases matter).
- To make a csv file, simply, put all your old words in the first column (A) in Microsoft Excel program and the new replacing words in the second column (B). Afterwards, save the file as csv (comma delimited) file.
- You can also invoke the save as window in Excel by simply hitting F12 button in Windows machines.