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:

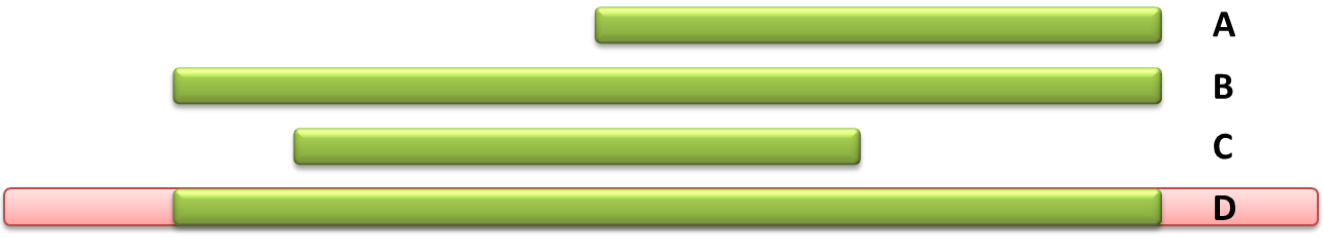
- Dereplication mode removes the 100% exact sequences and the shorter partial sequences.
- You can use it for one file or multiple files (with the same extension for which the program will ask you and in the same directory).
- You can use it on fasta file or fastq file.

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".

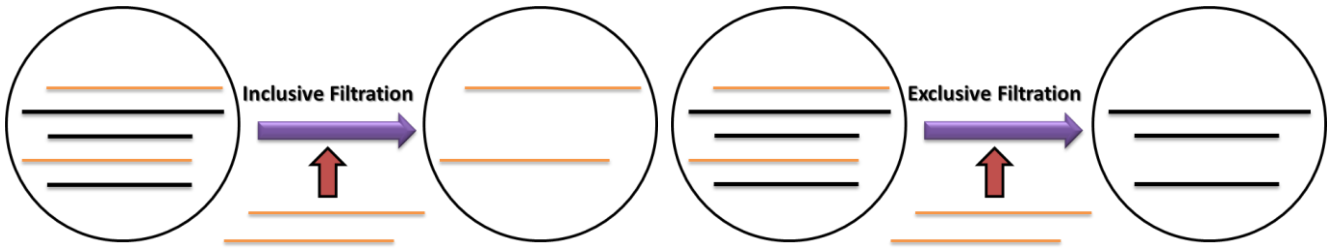
- You can set a minimum length option if you want to have a minimum length of your sequences.

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.
- The output of filteration 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 filteration is always exclusive.



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