filtered_nt

GW-Hive

Feb 04, 2022

SUMMARY OF THE PROTOCOL:

Filtered NT dataset is generated by excluding sequences from the whole nt file provided by NCBI, based on whether they have unwanted taxonomy names or any child taxonomy name of these unwanted ones. These unwanted taxonomy names are listed in the black list generated by two steps: (1) Getting all taxonomy names which contain the strings listed below (Step 3); (2) Getting all possible child taxonomy names of each of the taxonomy names from (1). For example, "other sequences" (taxId: 28384) is excluded with all its child taxonomy names including "artificial sequence", "vector", "synthetic", and so on.

We have chosen to apply the Creative Commons Attribution 3.0 Unsupported License to this version of the software.

Version	Downloadable Files	File Size	Release Notes	NCBI Download Date
Version 6.0	Filtered_NT v6.0	168G	Release Notes v6	July 2018
Version 5.0	Filtered_NT v5.0	131G	Release Notes v5.0	May 2017
Version 4.0	Filtered_NT v4.0	110G	Release Notes v4.0	July 2016

CHAPTER

ONE

SCRIPTS

CHAPTER

TWO

FILTER-NT MODULE

CHAPTER

THREE

STEP 1. DOWNLOAD THE WHOLE NT FILE

downloaded from: ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/ version: 5/21/2017 command: wget ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz gunzip nt.gz (42,439,338 rows)

CHAPTER

FOUR

STEP 2. DOWNLOAD THE TAXONOMY LIST

downloaded from: ftp://ftp.ncbi.nih.gov/pub/taxonomy/ version: 5/21/2017; 5/30/2017 command: wget ftp://ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid/*.gz gunzip *.gz wget ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz gunzip taxdump.tar.gz |tar -xvf location: /data/projects/targetdbs/downloads/

CHAPTER

FIVE

STEP 3. GENERATE BLACK LIST

- protocol: unwanted taxonomy names (scientific names) from names.dmp and all child taxonomy names of them, include: ['unclassified','unidentified','uncultured','unspecified','unknown','phage','vector'] ['environmental sample','artificial sequence','other sequence']

  There are two steps for generating the black list, first is to get all taxonomy names with the strings above, and then to get all child taxonomy names of them.

- script: /projects/targetdbs/scripts/get-parent-taxid-of-blacklist.py /projects/targetdbs/scripts/get-child-taxid-of-blacklist.py
- output: /data/projects/targetdbs/generated/blacklist-taxId.1.csv /data/projects/targetdbs/generated/blacklist-taxId.2.csv

  After generating blacklist-taxId.2.txt, use command line "sort -u" to delete duplicated records, and store them into: /data/projects/targetdbs/generated/blacklist-taxId.unique.csv

- QC script: /projects/targetdbs/scripts/compare-old-new-blacklist.py Compare the newly generated with the older version.

CHAPTER

SIX

STEP 4. CHECK THE COMPLETION OF TAXONOMY LIST (QC)

- protocol: First check if all seqAcs in nt file have taxIds from nucl_gb.accession2taxid file, and the ones do not have taxIds are checked in all other ac2taxid files.
- script: /projects/targetdbs/scripts/check-ac2taxid-completion-step1.py /projects/targetdbs/scripts/check-ac2taxid-completion-step2.py /projects/targetdbs/scripts/check-ac2taxid-completion-step3.py
- output: /data/projects/targetdbs/generated/logfile.step1.txt /data/projects/targetdbs/generated/logfile.step2.txt /data/projects/targetdbs/generated/logfile.step3.txt

  This step needs a lot of memory. Suggest to run on large machine. 123 records of PDB accessions have extra characters, fixed that in step3.py. However, 28 records are not in the files, search taxIds manually for them (/data/projects/targetdbs/generated/logfile.step3.manually.added.txt).

CHAPTER

SEVEN

STEP 5. GET THE SEQAC-TAXONOMY LIST

- protocol: Exclude those taxIds in the blacklist. And first get all seqAc-taxIds from nucl_gb.accession2taxid, and all of other ac2taxid files from both version 05/21/2017 and 05/30/2017.
- script: /projects/targetdbs/scripts/get-seqac2taxid.py
- output: /data/projects/targetdbs/generated/logfile.ac2taxid.list.txt
- QC step: All seqAcs in nt files are mapped to at least one taxId. The number of seqAcs in the list matches the one in nt file. SeqAcs with multiple taxIds are listed in: /data/projects/targetdbs/generated/seqAc-with-multiple-taxids.txt

CHAPTER

EIGHT

STEP 6. FILTERING NT FILE

- protocol: Remember to add those manually added ac2taxids.
- script: /projects/targetdbs/scripts/filter-nt.py
- output: /data/projects/targetdbs/generated/filtered_nt_Jun06-2017.fasta
- QC script: /projects/targetdbs/scripts/check-removed-seqacs-count.py