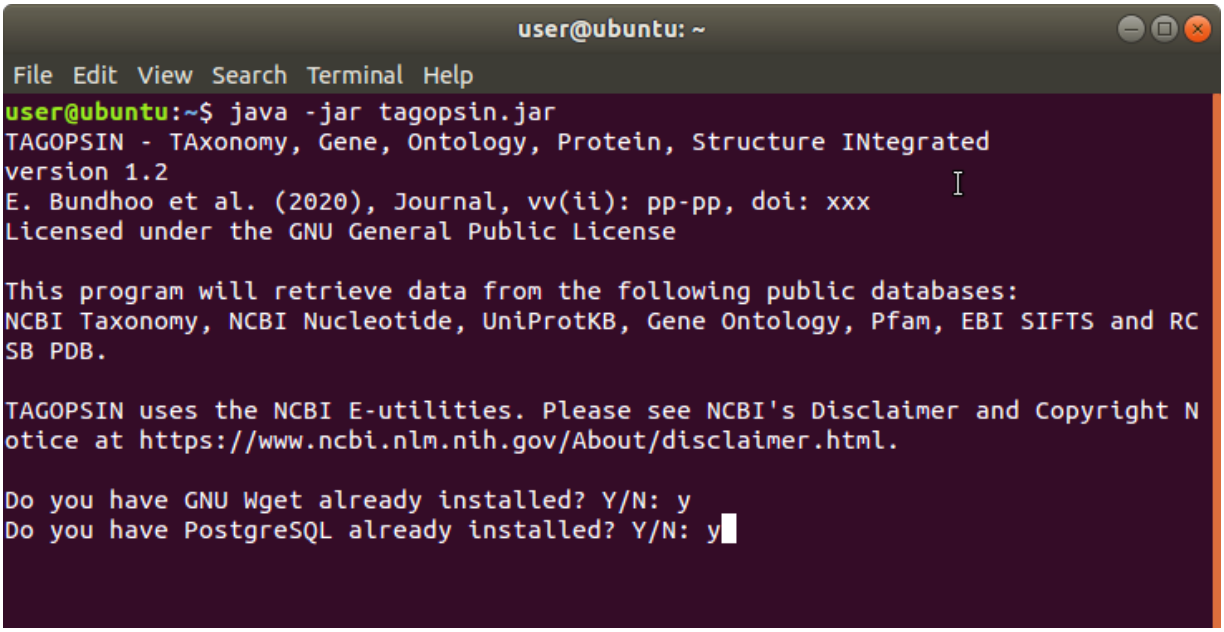
TAGOPSIN (version 1.2)

User Manual

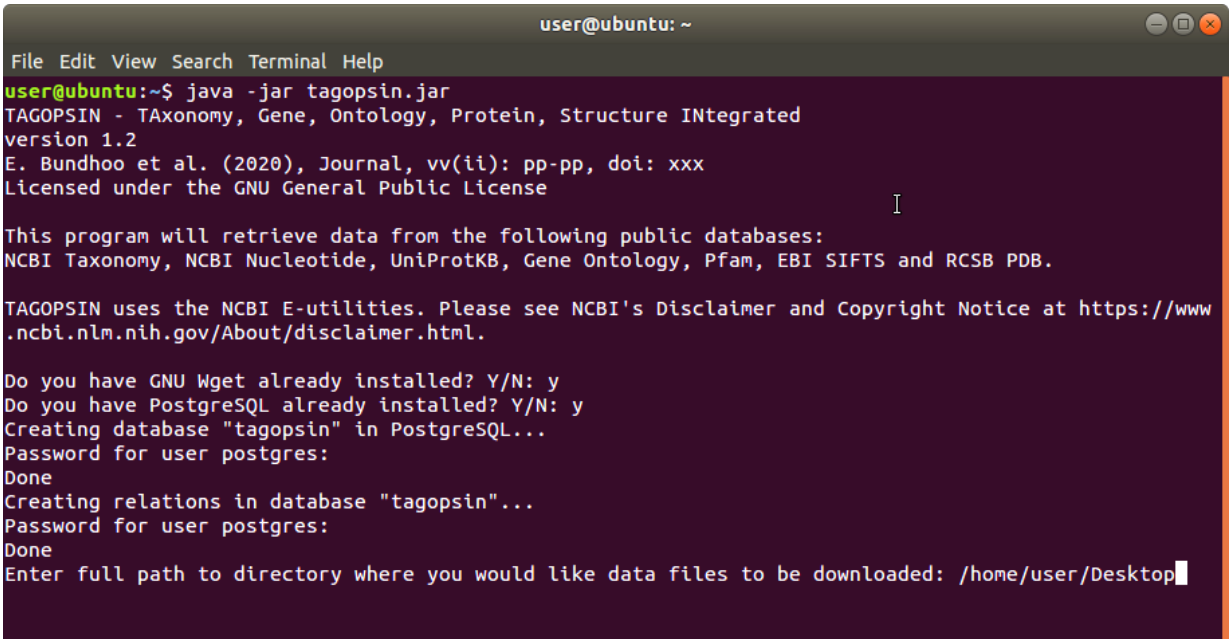
29 October 2020

Disclaimer: This step-by-step guide assumes that all dependencies (PostgreSQL, Wget and Java) are installed and working properly.

1. Go to the project home page on GitHub <https://github.com/ebundhoo/TAGOPSIN> and download the JAR file `tagopsin.jar`
2. `cd` to the directory where `tagopsin.jar` is saved
3. Type out the command `java –jar tagopsin.jar`
4. Once the program is running, you'll need to input some information
5. Answer "Y" or "y" (yes) to whether you have Wget and PostgreSQL installed



6. When running TAGOPSIN for the first time, you'll be prompted to enter a directory path. Type out the full path of a valid directory. All standard data files used by the program will be downloaded in this directory. In the example below, the desktop (`~/Desktop`) is specified as working directory.



7. TAGOPSIN will do a quick check of the validity of the directory path. So long as an invalid path is provided, you'll be prompted to re-enter this information.
8. After a valid directory is specified, TAGOPSIN will check all the files and sub-directories present in it. If any of the standard data files are absent, the program will download them successively.

```
Enter full path to directory where you would like data files to be downloaded: /home/user/Desktop
Checking "/home/user/Desktop"...
OK
Checking for presence of standard data files in "/home/user/Desktop"...
taxdump directory not found

At least one of the files required by TAGOPSIN is missing. Checking all files and directories in "/h
ome/user/Desktop"...
taxdump directory not found
Downloading ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdump.tar.gz...
```

9. If all files needed by the program are present in the working directory, it will proceed. The program will now establish a preliminary connection to the database "tagopsin." For this, it will request the URL location of the JDBC (Java Database Connectivity) driver, as well as the username and password to interface with PostgreSQL. Enter this information and press Return each time.

```
"/home/user/Desktop" was last used by TAGOPSIN
Checking for presence of standard data files in "/home/user/Desktop"...
names.dmp is present in "/home/user/Desktop/taxdump"
uniprot_sprot.dat and uniprot_sprot.fasta are present in "/home/user/Desktop/uniprot"
go-basic.obo is present in "/home/user/Desktop/gene_ontology"
Pfam-A.full.uniprot is present in "/home/user/Desktop/pfam"
All PDB files are present in "/home/user/Desktop/pdb"

All required data files are present. Proceeding now...
Connecting to database "tagopsin" in PostgreSQL...
Enter URL location of JDBC driver: jdbc:postgresql://localhost:5432/tagopsin
Enter username: postgres
Enter password:
Connection successful
```

10. You'll then be prompted to input the name and type of your organism of interest. Please refer to the README file available on the GitHub project home page for more details. Here, *Clostridium perfringens* is the organism of interest.
11. TAGOPSIN will now automatically retrieve data for this organism and insert them into the database "tagopsin" in PostgreSQL.

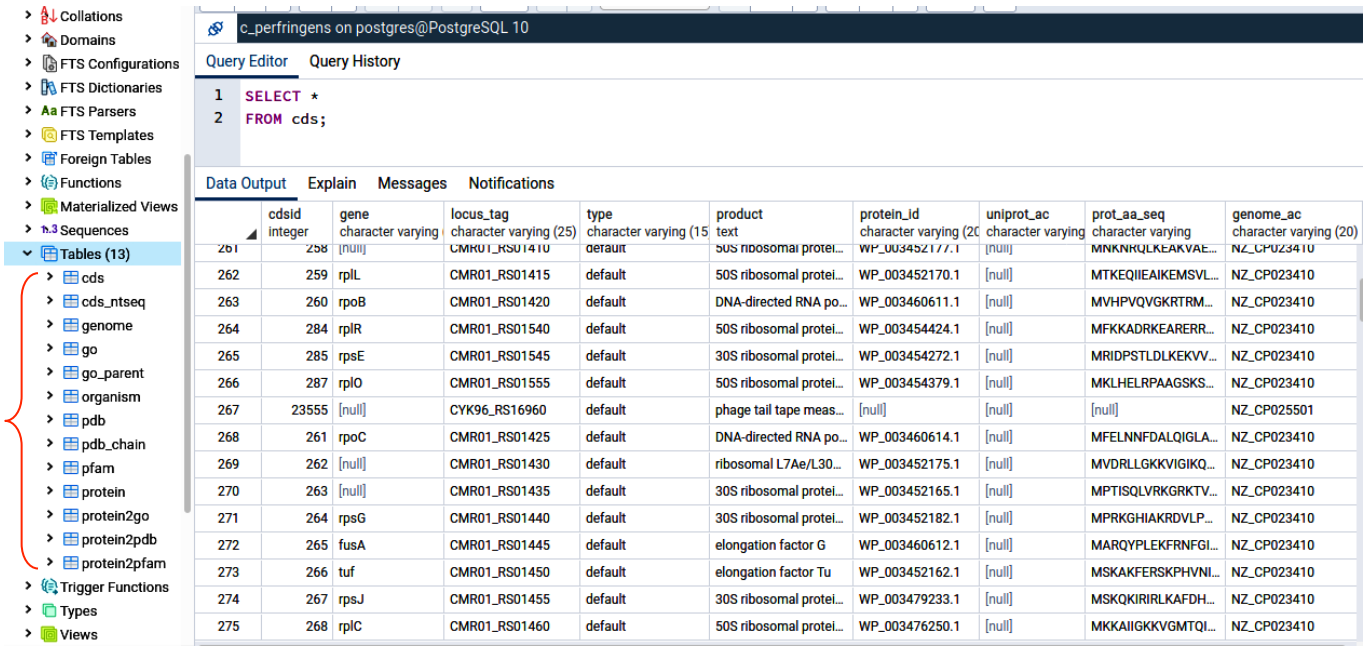
```
Depending on the objective of your project, organism name can be either Mycobacterium or Mycobacteri
um bovis for example.
Input name of organism of interest: Clostridium perfringens
Please specify whether this organism is eukaryotic (E), prokaryotic (P) or viral (V): p
Retrieving taxonomy IDs and scientific names from names.dmp...
```

12. Miscellaneous

(a) The tree view below shows how TAGOPSIN organises directories and files in the working directory (here `/home/user/Desktop`).

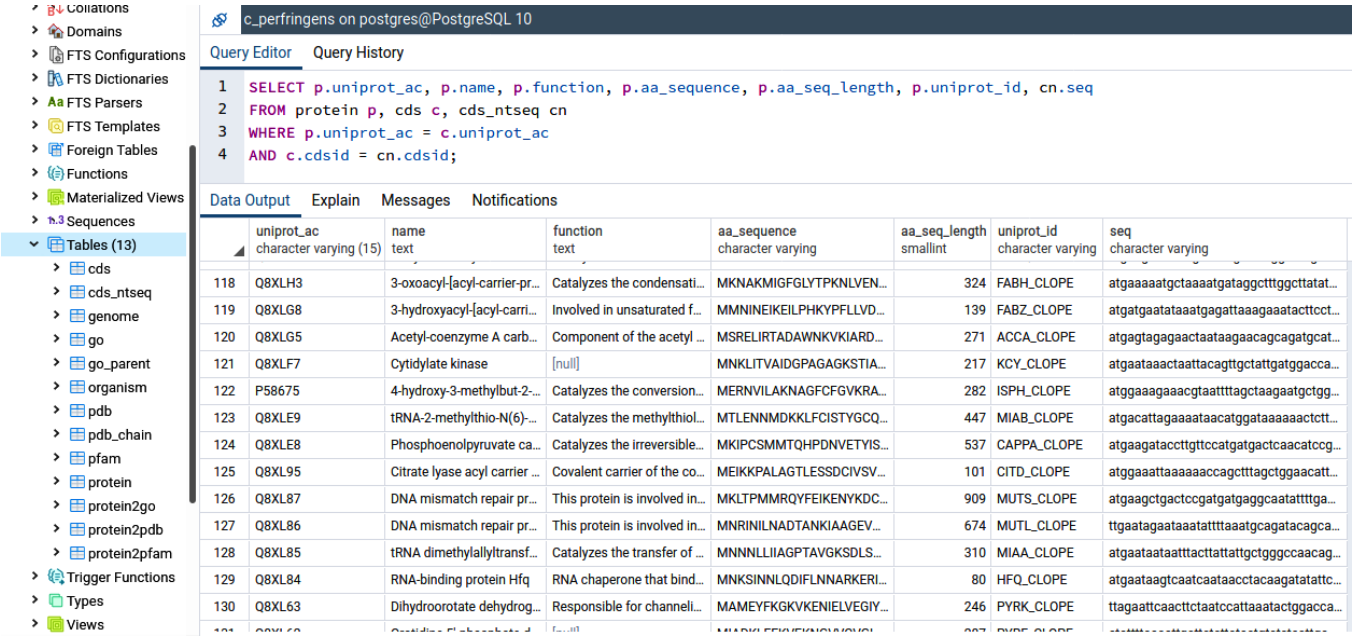
```
/home/user/Desktop/
├── genbank
│   ├── genomes
│   └── seqbins
├── gene_ontology
│   └── go-basic.obo
├── pdb
│   ├── entries.idx
│   ├── pdb_chain_taxonomy.tsv
│   ├── pdb_chain_uniprot.tsv
│   └── pdb_seqres.txt
├── pfam
│   └── Pfam-A.full.uniprot
├── taxdump
│   ├── citations.dmp
│   ├── delnodes.dmp
│   ├── division.dmp
│   ├── gc.prt
│   ├── gencode.dmp
│   ├── merged.dmp
│   ├── names.dmp
│   ├── nodes.dmp
│   ├── readme.txt
│   └── taxdump.tar.gz
└── uniprot
    ├── idmapping.dat
    ├── uniprot_sprot.dat
    └── uniprot_sprot.fasta
```

(b) Sample output of the program in PostgreSQL



	cdsid integer	gene character varying	locus_tag character varying (25)	type character varying (15	product text	protein_id character varying (20	uniprot_ac character varying	prot_aa_seq character varying	genome_ac character varying (20)
261	258	[null]	CMR01_RS01410	default	50S ribosomal protei...	WP_003452177.1	[null]	MNKNRQLKEAKVAE...	NZ_CP023410
262	259	rplL	CMR01_RS01415	default	50S ribosomal protei...	WP_003452170.1	[null]	MTKEQIIEAIKEMSVL...	NZ_CP023410
263	260	rpoB	CMR01_RS01420	default	DNA-directed RNA po...	WP_003460611.1	[null]	MVHPVQVGKRTRM...	NZ_CP023410
264	284	rplR	CMR01_RS01540	default	50S ribosomal protei...	WP_003454424.1	[null]	MFKKADRKEARERR...	NZ_CP023410
265	285	rpsE	CMR01_RS01545	default	30S ribosomal protei...	WP_003454272.1	[null]	MRIDPSTLDLKEKVV...	NZ_CP023410
266	287	rplO	CMR01_RS01555	default	50S ribosomal protei...	WP_003454379.1	[null]	MKLHELRPAAGSKS...	NZ_CP023410
267	23555	[null]	CYK96_RS16960	default	phage tail tape meas...	[null]	[null]	[null]	NZ_CP025501
268	261	rpoC	CMR01_RS01425	default	DNA-directed RNA po...	WP_003460614.1	[null]	MFELNNFDALQIGLA...	NZ_CP023410
269	262	[null]	CMR01_RS01430	default	ribosomal L7Ae/L30...	WP_003452175.1	[null]	MVDRLLGKKVIGIKQ...	NZ_CP023410
270	263	[null]	CMR01_RS01435	default	30S ribosomal protei...	WP_003452165.1	[null]	MPTISQLVRKGRKTV...	NZ_CP023410
271	264	rpsG	CMR01_RS01440	default	30S ribosomal protei...	WP_003452182.1	[null]	MPRKGHIAKRDVLP...	NZ_CP023410
272	265	fusA	CMR01_RS01445	default	elongation factor G	WP_003460612.1	[null]	MARQYPLEKFRNFGI...	NZ_CP023410
273	266	tuf	CMR01_RS01450	default	elongation factor Tu	WP_003452162.1	[null]	MSKAKFERSKPHVNI...	NZ_CP023410
274	267	rpsJ	CMR01_RS01455	default	30S ribosomal protei...	WP_003479233.1	[null]	MSKQKIRIRLKAFDH...	NZ_CP023410
275	268	rplC	CMR01_RS01460	default	50S ribosomal protei...	WP_003476250.1	[null]	MKKAIIGKKVGMTQI...	NZ_CP023410

I. Part of the GenBank dataset for *C. perfringens* available via the pgAdmin interface. The red bracket on the left shows the different relations built by TAGOPSIN in the local database "tagopsin."



	uniprot_ac character varying (15)	name text	function text	aa_sequence character varying	aa_seq_length smallint	uniprot_id character varying	seq character varying
118	Q8XLH3	3-oxoacyl-[acyl-carrier-pr...	Catalyzes the condensati...	MKNAKMIGFGLYTPKNLVEN...	324	FABH_CLOPE	atgaaaaatgctaaaatgataggctttggcttatat...
119	Q8XLG8	3-hydroxyacyl-[acyl-carri...	Involved in unsaturated f...	MMNINEIKEILPHKYPFLLVD...	139	FABZ_CLOPE	atgatgaatataaatgagattaaagaaatacttcct...
120	Q8XLG5	Acetyl-coenzyme A carb...	Component of the acetyl ...	MSRELIRTADAWNKVKIARD...	271	ACCA_CLOPE	atgagtagagaactaataagaacagcagatgcat...
121	Q8XLF7	Cytidylate kinase	[null]	MNKLITVAIDGPAGAGKSTIA...	217	KCY_CLOPE	atgaataaactaattacagttgctattgatggacca...
122	P58675	4-hydroxy-3-methylbut-2-...	Catalyzes the conversion...	MERNVILAKNAGFCFGVKRA...	282	ISPH_CLOPE	atggaaagaaacgtaattttagctaagaatgctgg...
123	Q8XLE9	tRNA-2-methylthio-N(6)-...	Catalyzes the methylthiol...	MTLENNMDKKLFCISTYGCQ...	447	MIAB_CLOPE	atgacattagaaaataacatggataaaaaactctt...
124	Q8XLE8	Phosphoenolpyruvate ca...	Catalyzes the irreversible...	MKIPCSMMTQHPDNVETYIS...	537	CAPPA_CLOPE	atgaagataccttgttccatgatgactcaacatccg...
125	Q8XL95	Citrate lyase acyl carrier ...	Covalent carrier of the co...	MEIKKPALAGTLESSDCIVSV...	101	CITD_CLOPE	atggaaattaaaaaaccagctttagctggaacatt...
126	Q8XL87	DNA mismatch repair pr...	This protein is involved in...	MKLTPMMRQYFEIKENYKDC...	909	MUTS_CLOPE	atgaagctgactccgatgatgaggcaatattttga...
127	Q8XL86	DNA mismatch repair pr...	This protein is involved in...	MNRINILNADTANKIAAGEV...	674	MUTL_CLOPE	ttgaatagaataaatattttaaatgcagatacagca...
128	Q8XL85	tRNA dimethylallyltransf...	Catalyzes the transfer of ...	MNNNLLIIAGPTAVGKSDLS...	310	MIAA_CLOPE	atgaataataatttacttattattgctgggccaacag...
129	Q8XL84	RNA-binding protein Hfq	RNA chaperone that bind...	MNKSINNLQDIFLNNARKERI...	80	HFQ_CLOPE	atgaataagtcaatcaataacctacaagatatattc...
130	Q8XL63	Dihydroorotate dehydrog...	Responsible for channeli...	MAMEYFKGKVKENIELVEGIY...	246	PYRK_CLOPE	ttagaattcaacttctaatccattaaatactggacca...

II. Part of the UniProt dataset for *C. perfringens* available via the pgAdmin interface. The last column “seq” contains the coding sequence from GenBank for each individual UniProt entry.