**Virulence Data Information**

The file pickles\_and\_pngs.tar.gz contains the following:

**Genera: Campylobacter, Escherichia, Listeria, Salmonella, Shigella**

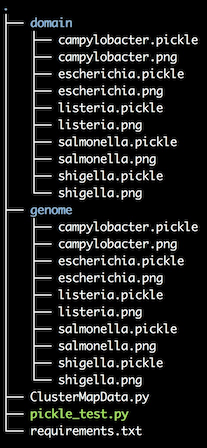
**domain** - contains pickle and png files from the analysis of pivot domain architecture vs neighbor domain architecture

**genome** - contains pickle and png files from the analysis of genome vs domain architecture

**ClusterMapData.py** - Python class file that needs to be imported in order to create the deserialized version of the pickled object.

**requirements.txt** - dependencies

**pickle\_test.py** - a small python program that loads each pickle and constructs a ClusterMapData instance and outputs some info about it.



IBM collaborators used a python 3.7 virtual environment. To create run:

* python -m venv venv
* . ./venv/bin/activate
* pip install -r requirements.txt
* ./pickle\_test.py

Reference data such as domain architectures, protein names, virulence data, etc. can be found in the **csv\_files.tar.gz** file. This file contains the following:

**domain\_architecture.csv** - maps unique IPR sets to a common UID key.

* Headers:
  + DOMAIN\_ARCHITECTURE\_UID\_KEY
  + DOMAIN\_ARCHITECTURE

**ipr.csv** - provides name, description and type for each IPR accession in domain\_architecture.csv

* Headers:
  + IPR\_ACCESSION
  + NAME
  + DESCRIPTION
  + TYPE

**protein\_domain.csv** - cross reference between protein and domain architecture.

* Headers:
  + PROTEIN\_UID\_KEY
  + DOMAIN\_ARCHITECTURE\_UID\_KEY

**protein\_name.csv** - maps OMXWare protein full names to UID key.

* Headers:
  + PROTEIN\_UID\_KEY
  + PROTEIN\_FULLNAME

**protein\_virulence.csv** - maps VFDB (<http://www.mgc.ac.cn/VFs/main.htm>) reference to PROTEIN\_UID\_KEY

* Headers:
  + PROTEIN\_UID\_KEY,
  + SHORT\_NAME
  + FULL\_NAME
  + VIRULENCE\_FACTOR
  + GENUS
  + SPECIES
  + STRAIN