**Virus Target Set Overlap Analysis**

**Overview**

This analysis quantifies the overlap between target sets of virus-interacting human proteins for multiple viruses. The goal is to assess the similarity between human protein targets of different viruses using two standard metrics: the Overlap Coefficient (C) and the Jaccard Index (J).

## Input Files

1. **Virus Pairs File** (Network\_Separation.xlsx):
   * An Excel file containing virus pairs to compare.
   * Example structure:

| **Virus\_A** | **Virus\_B** |
| --- | --- |
| Dengue\_virus\_type\_2\_(DENV\_2) | Hendra\_virus\_(HeV) |
| Dengue\_virus\_type\_2\_(DENV\_2) | Hepatitis\_C\_virus\_(HCV) |
| ... | ... |

1. **Virus Gene Lists** (<Virus\_Name>.txt):
   * Plain text files containing human protein targets for each virus.
   * Each line represents a single protein ID.
   * Example: Dengue\_virus\_type\_2\_(DENV\_2).txt, Hendra\_virus\_(HeV).txt, Hepatitis\_C\_virus\_(HCV).txt

**Note:** Place the Excel file and all virus .txt files in the same folder as the Python script

## Metrics

1. **Overlap Coefficient (C):**

*C = |AՈB|/min(|A|,|B|)*

1. **Jaccard Index (J):**

*J = |AՈB|/|AՍB|*

Both metrics range from 0 to 1, where values of J and C equal to 0 indicate no common virus-interacting human proteins. Partial overlap is indicated when 0 < J < 1 and 0 < C < 1, while J=1 and C=1 indicate that A and B are identical. A Jaccard Index of J < 1 and an Overlap Coefficient C = 1 suggest that A is a complete subset of B.

## Python / Jupyter Notebook Script

* File name: Jaccard\_Index\_and\_Overlap\_Coefficient.ipynb
* The script/notebook performs the following steps:
  1. Reads virus pairs from Network\_Separation.xlsx.
  2. Loads each virus gene list from the corresponding .txt files.
  3. Computes the **Overlap Coefficient (C)** and **Jaccard Index (J)** for each virus pair.
  4. Saves the results to virus\_pairs\_results.csv.

## **Key features:**

* Handles missing .txt files gracefully.
* Works with files in the same folder (no absolute paths required).
* Outputs results in an easy-to-read CSV format.

## Output

The output file, virus\_pairs\_results.csv, contains the overlap results for all virus pairs:

| **Virus\_A** | **Virus\_B** | **Overlap Coefficient** | **Jaccard Index** |
| --- | --- | --- | --- |
| Dengue\_virus\_type\_2\_(DENV\_2) | Hendra\_virus\_(HeV) | 0.17 | 0.007 |
| Dengue\_virus\_type\_2\_(DENV\_2) | Hepatitis\_C\_virus\_(HCV) | 0.26 | 0.14 |
| ... | ... | ... | ... |

## Notes

* This analysis can be extended to **larger virus datasets**.
* It provides a **quantitative comparison** of virus-interacting human proteins.
* Combined with functional enrichment or hypergeometric tests, it can reveal **significant overlaps and unique targets**.

## Running in Jupyter

1. Open Jaccard\_Index\_and\_Overlap\_Coefficient.ipynb in Jupyter Notebook or Jupyter Lab.
2. Ensure Network\_Separation.xlsx and all virus .txt files are in the same folder.
3. Run all cells sequentially.
4. The results will be saved as virus\_pairs\_results.csv in the same folder.