Guideline for Reproducing the Antigenic Evolution Results

This document describes how to reproduce the antigenic evolution analyses presented in **"Unraveling the mechanism behind the probable extinction of the B/Yamagata lineage of influenza B viruses" (Nature Communications, 2025)**.

The workflow consists of six major components:

1. Data processing
2. Feature calculation
3. Model construction
4. Antigenic cluster identification
5. Identification of antigenic-cluster–transition–associated sites
6. Model comparison with published models

Please ensure that all input and output paths in the scripts are correctly configured before execution. For illustration, we describe the workflow for **B/Victoria**; the **B/Yamagata** pipeline follows the same structure.

1. Data Processing

Directory: ./BV/data_processing

This folder contains scripts for preprocessing HA sequences and hemagglutination inhibition (HI) assay data.

2. Feature Calculation

Directory: ./BV/features

This module computes four classes of antigenically relevant features for each sequence pair:

- **Epitope scoring**: BV_pssm.py
- **Physicochemical property scoring**: BV_aaindex.py
- **Glycosylation scoring**: BV_Ngly.py
- **Receptor-binding site (RBS) scoring**: BV_RBS.py

These features serve as inputs to the antigenic prediction model.

3. Model Construction

Directory: ./BV/model_construction

This folder contains the scripts for building the antigenic prediction model using the feature sets above. Models are trained to predict antigenic relationship between strain pairs, following the procedures described in the manuscript.

4. Antigenic Cluster Identification

Directory: ./BV/antigenic_clusters

This component identifies antigenic clusters based on predicted antigenic relationship and Markov clustering (MCL).

Key scripts:

- **BV_crosstable.py**
  Generates all pairwise combinations of non-redundant sequences. Because the full pairwise matrix is very large (millions of strain pairs), we do not

provide it directly. Users can reproduce the file using the provided non-redundant sequence dataset (BV_sample_dropoutlier.csv in data/sequence/).

- **features/**
  Contains scripts to calculate the full feature matrix for all generated strain pairs.
- **batch_mcl.py**
  Defines the function for executing MCL over a specified range of inflation parameters.
- **Mean_Cluster_Size.py**
  Computes mean antigenic cluster size for each inflation value.
  (Note: This metric was explored but not used for selecting the optimal inflation parameter.)
- **MCL_function.py**
  Computes modularity (which were used to identify the optimal inflation parameter), hemisphere-level and continent-level cluster proportions.
- **BV_MCL_command.py**
  A wrapper script that runs the MCL workflow by calling the three MCL-related modules above.

5. Identification of Antigenic Cluster–Transition–Associated Sites
Directory: ./BV/determining_sites
This folder contains scripts to identify HA1 amino acid positions associated with transitions between antigenic clusters. Shannon entropy and information gain are used to quantify the importance of each site.

6. Model Comparison with Published Models
Directory: ./model_compare
This module includes:
Scripts for comparing the performance of our antigenic prediction model with two previously published antigenic models.