Autoencoder Package Instructions

Data PreProcessing

1. Run NCBI_Extract.py to obtain sequences from NCBI → outputs .fasta
2. If gathering entire sequences:
    a. Use file_separator.py to split each identifier into its own .txt file
    b. Run Complete_Sequence_Retrieval_And_Cluster.py to obtain sequence by start/end motif → outputs Dataframe and fasta file
    c. OR run Protein_Extract to extract by protein instead
3. If gathering specific proteins: skip to step 2 in Experiment
4. Model accepts sequences where file name is in column 1, and there is 1 amino acid position in each cell up to 3000 positions.
5. Model does not handle insertions and deletions well, will have to filter out at the moment.

Experiment

1. From raw files, merge to one .txt using Merge_All_Text_To_One.py → outputs a merged_output.fasta file (NONPROCESSED)
2. Use clustalo -i and -o to align i.e. (clustalo -i /home/blim/Documents/Summer2023/DENV2.fasta -o /home/blim/Documents/Summer2023/DENV2_aligned.fasta) → outputs a .fasta file that is aligned
3. With aligned file. Run it through Convert_Fasta_to_Excel_2.py → outputs a .output.xlsx file with each aa in its own column
4. Run through AE. first part specifies sequence starting from residue number, and then removes all X/J values. Will a
5. lso replace "-" values with the mode if necessary.
    a. After autoencoder, will output a top10 mutation df.

DENV2_aligned.fast -> All_DENV2_aligned.xlsx -> DENV2_ENV.xlsx -> topmutations

If you want to study all serotypes together:

1. Run through Merge_All_Excel_to_one.py. Make sure all DENV(X)_ENV.xlsx are in same folder. Output All_DENV_ENV.xlsx
2. Run Convert_Excel_to_Fasta.py → outputs output.fasta (All_DENV_ENV.fasta)
3. Align with clustal
4. Probably better way to do it, but i removed all nan values with Ctrl+F since they're only at the end. Also, removed all "-" values that were at the end
5. Run Convert_Fasta_to_Excel_2.py → output All_DENV_ENV_aligned.xlsx
6. Replaced "-" values with the mode in first cell of DENV_AE.
7. Run through AE model

Analysis

1. Run Mutant_finder.py → outputs .xlsx that identifies all mutations within data sample when compared to the mode at the position. Verify if mutations exist within our dataset or predicting new mutations.
2. Add pre-processed amino acid position to global position in autoencoder model analysis section to more easily search the amino acid residue.
3. Run heatmap to analyze amino acid probability at each position.
4. Check MSE and decoded vs reconstructed data for accuracy.

Finished Codes
Convert_Excel_to_Fasta.py
Convert_Fasta_to_Excel.py
Merge_All_Excel_to_one.py
Mutant_finder.py
Complete_Sequence_Retrieval_And_Cluster.py