Visualising SARS-CoV-2 Evolutionary Space Using Protein Language Models

Brendonas Stakauskas

Institute of Data Science and Digital Technologies, University of Vilnius

Introduction

Large Language Models (LLMs) based on Transformer architecture have been adapted for protein sequence analysis, predicting structures and functions directly from sequences. These models have shown significant success in understanding the dynamics of viral mutations. Phylogenetic trees represent evolutionary relationships among various biological species based on their genetic information. This work combines LLM-based protein embeddings with phylogenetic trees to visualize and analyze SARS-CoV-2 mutations.

Dataset and Processing

41,387 SARS-CoV-2 sequences from Lithuania (Feb 2020 - Mar 2023) were sourced from GISAID. Only the spike glycoprotein gene was selected for embedding, as this is the main target for immune response. This gene sequence consists of 1273 amino acids.

Data preprocessing

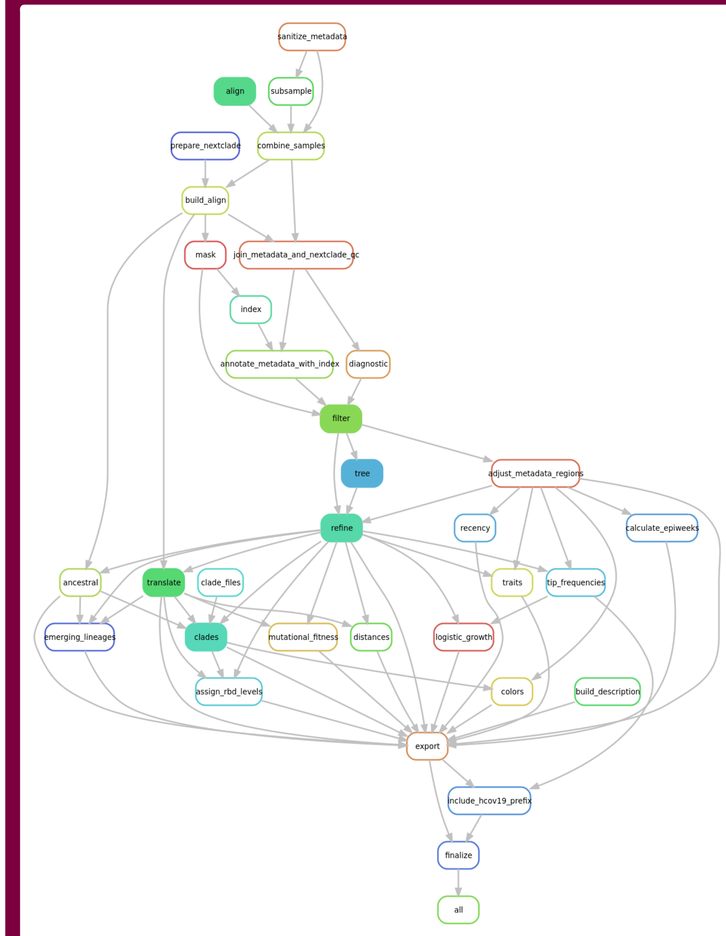


Figure 1: Data processing workflow from Nextstrain. The most critical steps for this analysis are highlighted.

Phylogenetic trees were built using FastTree. Embeddings were generated using ESM-1b and ESM-2 models. Nextstrain workflow was used for data preprocessing. Data workflow is visualized in Figure 1. Data embedding pipeline is shown in Figure 2.

Embedding

ESM class models were tested, namely ESM-1b and ESM-2 consisting of 650 million and 3 billion parameters respectively with the embedding dimensions 1280 and 2560. Embeddings are built for every token in sequence. As the data size is huge we used the mean of outputs from the model.

Data embedding

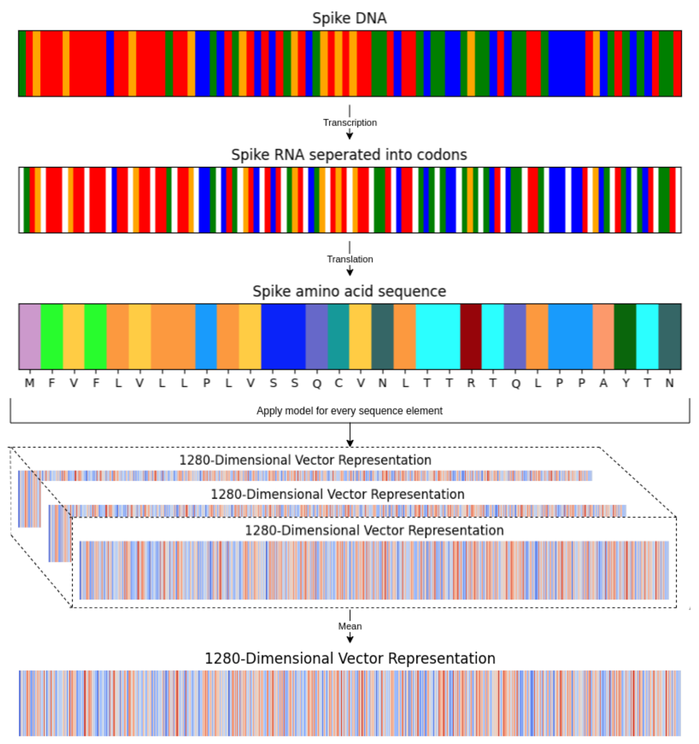


Figure 2: ESM-2B model embeddings pipeline for outputting 1280-dimensional data.

Embedded data vectors were visualized in 2D space using MDS and TSNE algorithms, using either sequencing date or assigned pangolin lineage as a label (Figure 4).

Phylogenetic paths

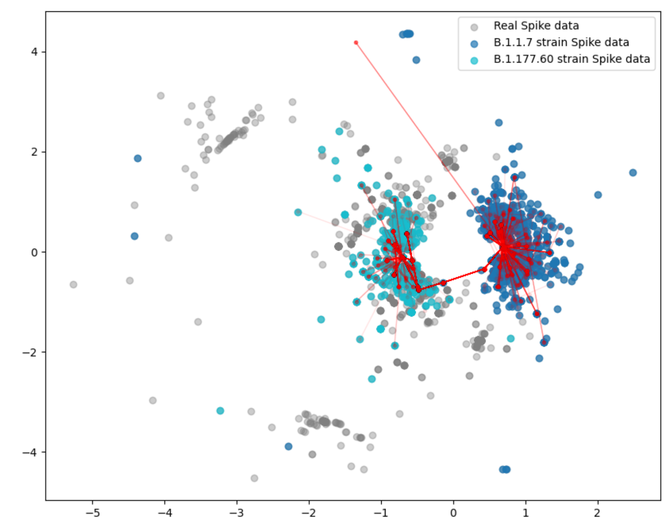
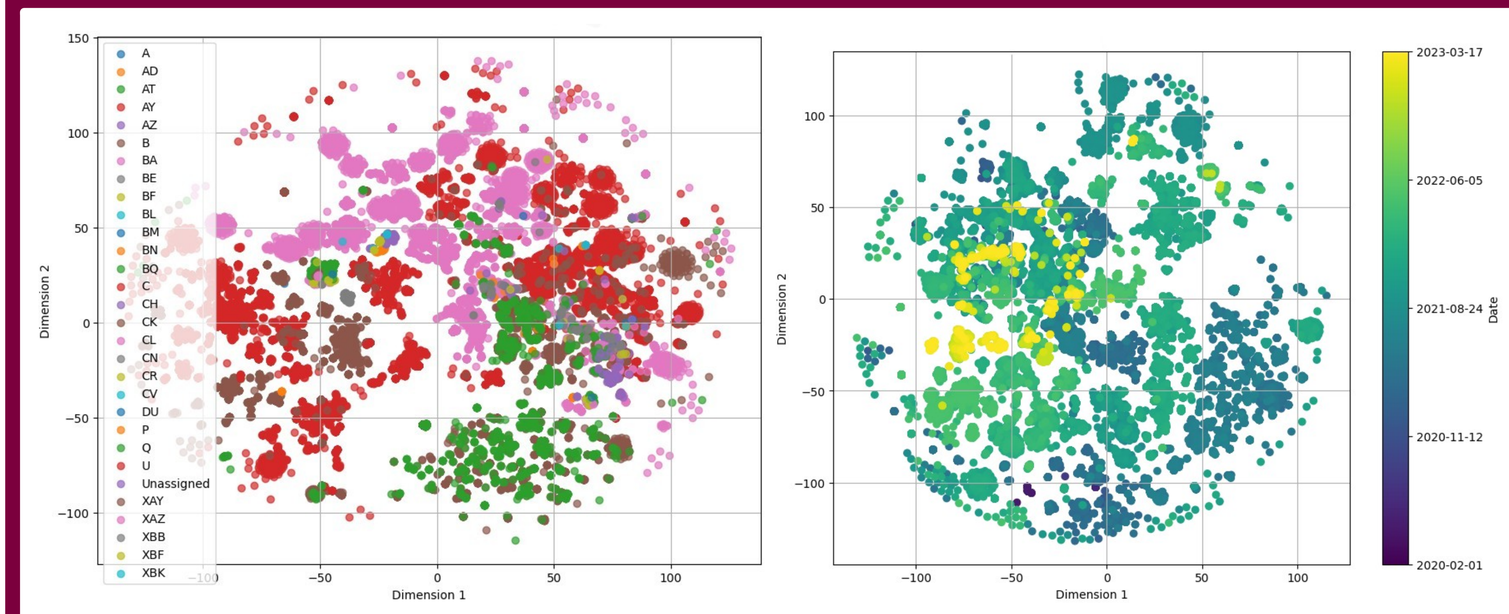


Figure 3: MDS transformation of nodes inferred by phylogenetic tree algorithm.

The internal nodes, inferred from phylogenetic tree, were also used. Firstly the bigger viral subsets were separated according to their pango lineage. Most common strains were chosen. Then for every point in those strains phylogenetic path was drawn(Figure 3). This was done in order to inspect the mutational path in the created projection of data.

2D visualization



TSNE by lineage

TSNE by date

Figure 4: TSNE transformations of ESM-1b and ESM-2 data embeddings.

Results and discussion

In this work only the spike part of the sequence was used. This led to phylogenetic paths being short, lacking information in the mutational space. Future research should tackle this problem by expanding the dataset to whole virus sequence.