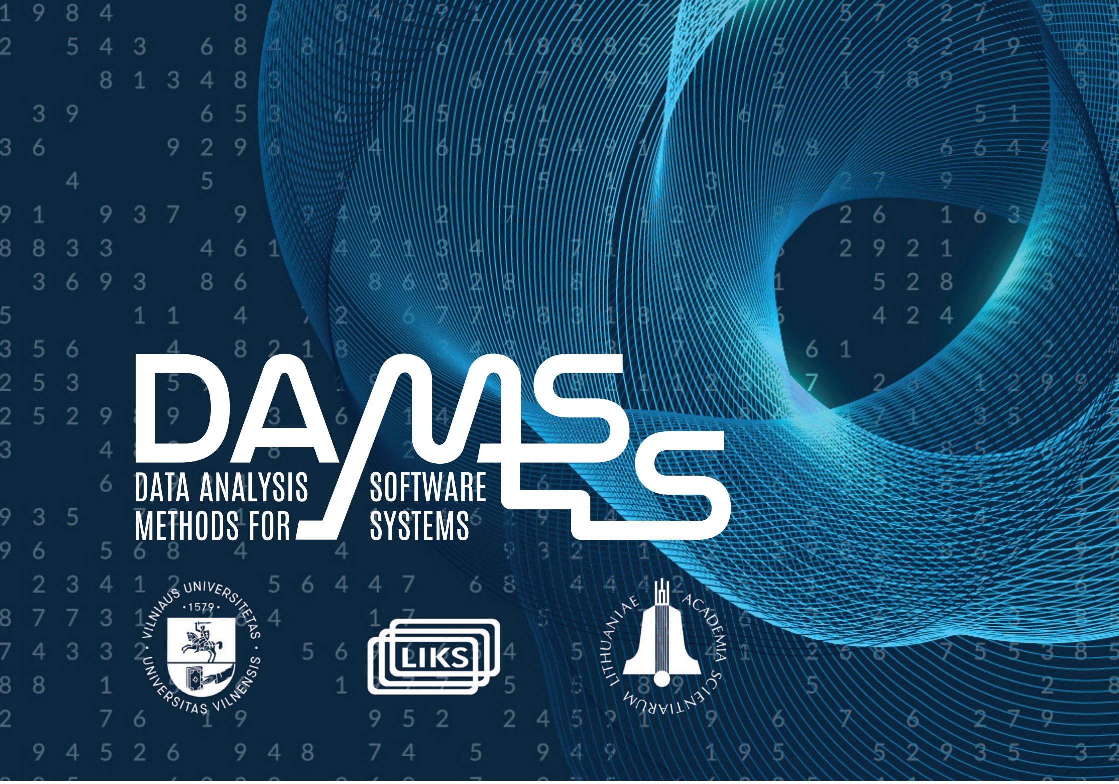
DAMSS
DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS
VILNIAUS UNIVERSITETAS •1579• UNIVERSITAS VILNENSIS
LIKS
ACADEMIA SCIENTIARUM LITHUANIAE

15th Conference on

DATA ANALYSIS METHODS for Software Systems

November 28–30, 2024

Druskininkai, Lithuania, Hotel "Europa Royale"
https://www.mii.lt/DAMSS

Co-Chairmen:

Prof. **Gintautas Dzemyda** (Vilnius University, Lithuanian Academy of Sciences)
Dr. **Saulius Maskeliūnas** (Lithuanian Computer Society)

Programme Committee:

Dr. **Jolita Bernatavičienė** (Lithuania)
Prof. **Juris Borzovs** (Latvia)
Prof. **Janis Grundspenkis** (Latvia)
Prof. **Janusz Kacprzyk** (Poland)
Prof. **Ignacy Kaliszewski** (Poland)
Prof. **Bożena Kostek** (Poland)
Prof. **Tomas Krilavičius** (Lithuania)
Prof. **Olga Kurasova** (Lithuania)
Assoc. Prof. **Tatiana Tchemisova** (Portugal)
Assoc. Prof. **Gintautas Tamulevičius** (Lithuania)
Prof. **Julius Žilinskas** (Lithuania)

Organizing Committee:

Dr. **Jolita Bernatavičienė**
Prof. **Olga Kurasova**
Assoc. Prof. **Viktor Medvedev**
Laima Paliulionienė
Assoc. Prof. **Martynas Sabaliauskas**
Prof. **Povilas Treigys**

Contacts:

Dr. Jolita Bernatavičienė
jolita.bernataviciene@mif.vu.lt
Tel. (+370 5) 2109 315
Prof. Olga Kurasova
olga.kurasova@mif.vu.lt

https://doi.org/10.15388/DAMSS.15.2024

ISBN 978-609-07-1112-5 (digital PDF)

Visualising SARS-CoV-2 Phylogenetic Relationships Using Protein Language Models

Brendonas Stakauskas, Virginijus Marcinkevičius

Institute of Data Science and Digital Technologies
Vilnius University

brendonas.stakauskas@mif.stud.vu.lt

Recent advancements have demonstrated the efficacy of Large Language Models (LLMs) based on Transformer architecture for natural language processing tasks, which have been successfully adapted for protein sequence data. Transformer-based models pre-trained on extensive protein databases can predict structures, functions, and other properties from protein sequences alone. These models have shown significant success in understanding the dynamics of viral mutations.

Viruses constantly mutate, and the changes that appear in viral code can lead to different reactions in the host body. Constant changes complicate vaccination, as the new strains can be immune to older vaccines. To monitor the dynamics of mutational changes phylogenetic tree building algorithms are used. Phylogenetic trees represent evolutionary relationships among various biological species based on their genetic information. In this work, we study similarity-based methods for phylogenetic relationship (evolutionary path) visualisation using protein sequence embeddings from protein language models.

The data is gathered from the GISAID Data Science Initiative. This data provider is popular among laboratories gathering sequential data and researchers that study these datasets.

Our used dataset consisted of 41387 SARS-CoV-2 sequences that were collected in Lithuania from February 2020 to March 2023. The dataset spans a comprehensive three-year period (2020-2023), covering key phases of the pandemic in Lithuania. Studying a geographically confined and heavily regulated population may result in lower viral diversity, while the extended timeframe allows us to capture the full extent of the virus's evolutionary dynamics over time. Data processing was applied using next strain workflow (available on the next strain GitHub page). Data processing included phylogenetic tree building, which was done using IQ-Tree and TreeTime algorithms. Data was filtered to only include virus samples that were collected from human hosts.

Embeddings were built using ESM class models, namely ESM-1b and ESM-2 consisting of 650 million and 3 billion parameters, respectively. Embeddings are built for every token in sequential data of protein. As the data size is huge, we used the mean of outputs from the model. Script for embedding extraction can be found in ESM GitHub repository.

For visualisation purposes the data dimension was shrunk using dimensionality reduction techniques such as TSNE and UMAP. The data points were grouped according to the sequence metadata such as sequencing date or their assigned pango lineage. Although some patterns emerge in produced data plots, the overlapping exists. Internal nodes, acquired from tree-building algorithms, were used in viral evolution visualisations. The phylogenetic tree paths for this dataset were deep thus making it hard to make sense of the information present in the tree. Evolutionary steps acquired from the tree are sparse, and evolutions appear on the different proteins in the SARS-CoV-2 virus. We investigate techniques like node and protein joining to build a dataset from the protein embeddings and phylogenetic tree data, allowing for the visualisation of viral mutation dynamics and the similarity between individual virus samples. The goal of this study is to investigate the consistency of these relationships.