

Visualization of the Evolutionary Path: an Influenza Case Study

Majid Forghani^{1,2}, Pavel Vasev², Edward Ramsay³ and Alexander Bersenev^{1,2}

¹*N.N. Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of the Russian Academy of Sciences (IMM UB RAS), 16 S.Kovalevskaya St., Yekaterinburg, 620108, Russia*

²*Ural Federal University, 51 Lenina Ave., Yekaterinburg, 620075, Russia*

³*Smorodintsev Research Institute of Influenza, Russian Ministry of Health, 15/17 Ulitsa Professor Popova, St. Petersburg, 197376, Russia*

Abstract

Visualization of viral evolution is one of the essential tasks in bioinformatics, through which virologists characterize a virus. The fundamental visualization tool for such a task is constructing a dendrogram, also called the phylogenetic tree. In this paper, we propose the visualization and characterization of the evolutionary path, starting from the root to isolated virus in the leaf of the phylogenetic tree. The suggested approach constructs the sequences of inner nodes (ancestors) within the phylogenetic tree and uses one-hot-encoding to represent the genetic sequence in a binary format. By employing embedding methods, such as multi-dimensional scaling, we project the path into 2D and 3D spaces. The final visualization demonstrates the dynamic of viral evolution locally (for an individual strain) and globally (for all isolated viruses). The results suggest applications of our approach in: detecting earlier changes in the characteristics of strains; exploring emerging novel strains; modeling antigenic evolution; and study of evolution dynamics. All of these potential applications are critical in the fight against viruses.

Keywords

Visualization, Influenza, Evolutionary path, Evolution, H3N2

1. Introduction

Viruses are an integral part of human life. Some viruses, e.g., influenza, hepatitis, and HIV, pose a severe threat to public health. Viruses affect not only public health, but also have serious consequences for the economy. For this reason, the activity of viruses, especially the influenza virus, is continuously monitored by the World Health Organization to study their evolution and to combat them [1]. Beyond the influenza virus, the recently emergent COVID-19 pandemic has, once again, reminded us of the importance of studying evolution and its hidden mechanisms. The study of evolution, and characterization of causative agents, are crucial factors in vaccine production. Evolution causes the virus to alter the structure and properties of antigens, through gradual accumulation of genetic mutations, leading to escape from immune responses [2]. This

GraphiCon 2021: 31st International Conference on Computer Graphics and Vision, September 27–30, 2021, Nizhny Novgorod, Russia

✉ forghani@imm.uran.ru (M. Forghani); vasev@imm.uran.ru (P. Vasev); WarmSunnyDay@mail.ru (E. Ramsay); bay@hackerdom.ru (A. Bersenev)

🆔 0000-0002-9443-3610 (M. Forghani); 0000-0003-3854-0670 (P. Vasev); 0000-0001-7086-5825 (E. Ramsay); 0000-0001-5843-5224 (A. Bersenev)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

leads to loss of vaccine efficacy and to revision requirements. Therefore, studies of antigen evolution are an essential part of the strategy in fighting viruses.

Visualization aims at providing a new scientific understanding, or viewpoint, that allows the researcher to better observe, explore, or receive insight from data [3]. A phylogenetic tree is a representation of evolutionary history. It is one of the most fundamental data structures in biology, showing a compact form of evolution through similarities and differences between genetic sequences [4]. In fact, the tree data structure transforms the complex evolutionary relationships between species into a graphic, human-readable representation [5].

Mainly, the phylogenetic tree is constructed through two steps: computing a distance matrix; and inferring a tree topology from the matrix. Various models, such as Kimura-80 (K80) [6], have been developed previously to compute the distance matrix from a set of genetic sequences. Construction of a tree structure relies on clustering the species based on their distances. The clustering can be carried out by different algorithms, including classical methods, such as neighbor-joining [7], Fitch–Margoliash [8], etc. The phylogenetic tree is a branching diagram, which can be represented in a variety of forms, e.g., rooted/unrooted, circular tree, cladogram [9], phylogram [10], and coral of life [11].

The phylogenetic tree can be directly used in modeling antigenic evolution. Successful examples of such applications have been presented [12, 13]. The idea relies on employing the relationship between the paths on the tree (which connects the pair of reference and test viruses) and their associated hemagglutinin inhibition assay data for modeling. Another example of the application of visualization for modeling was proposed by Ito et al. [14]. Their approach relies on predicting the evolutionary direction and identifying the viruses near the direction that can spread in the upcoming season. The key idea of their approach is constructing a three-dimensional map from the hemagglutinin sequences of the H3N2 subtype. The map is created by computing the genetic distance matrix and applying embedding algorithms, such as multidimensional scaling (MDS) [15], to project the distance matrix space of the viruses into a 3D space. Their research indicates that viruses located near the evolutionary direction have great vaccine potential and need to be the subject of further analysis.

As mentioned earlier, our goal is to visualize the evolutionary path of a strain over time. The main idea of this paper is inspired by Rubik’s cube solving algorithms [16]. A solution path is visualized from a random initial state to the final solution by the one-hot-encoding and t-SNE method [17]. The evolutionary path is a path from the root to a leaf of the phylogenetic tree. Implementing the visualization of such a path requires access to the tree’s genetic sequence of internal nodes. This can be handled by reconstructing the ancestral sequences. Like a Rubik’s cube visualization, we use the reverse evolutionary path, such that the root will be the final solution. Therefore, each solution path starts from a leaf and ends at the root. By matching the root coordinates, the final form of viral evolution for a tree is achieved.

Our contributions in this paper mainly focus on establishing a novel representation of the evolutionary path, which can further be employed in other studies, such as antigenic evolution modeling. The rest of the paper is organized as follows. Section 2 explains the proposed method in more detail. Section 3 is devoted to experiment setup and results. Finally, the conclusion is given in Section 4.

2. Methodology

Figure 1 illustrates the overall schema of our approach. The schema includes two main steps: performing the phylogenetic analysis; and constructing the visualization. Each step is described in more detail in the following sections.

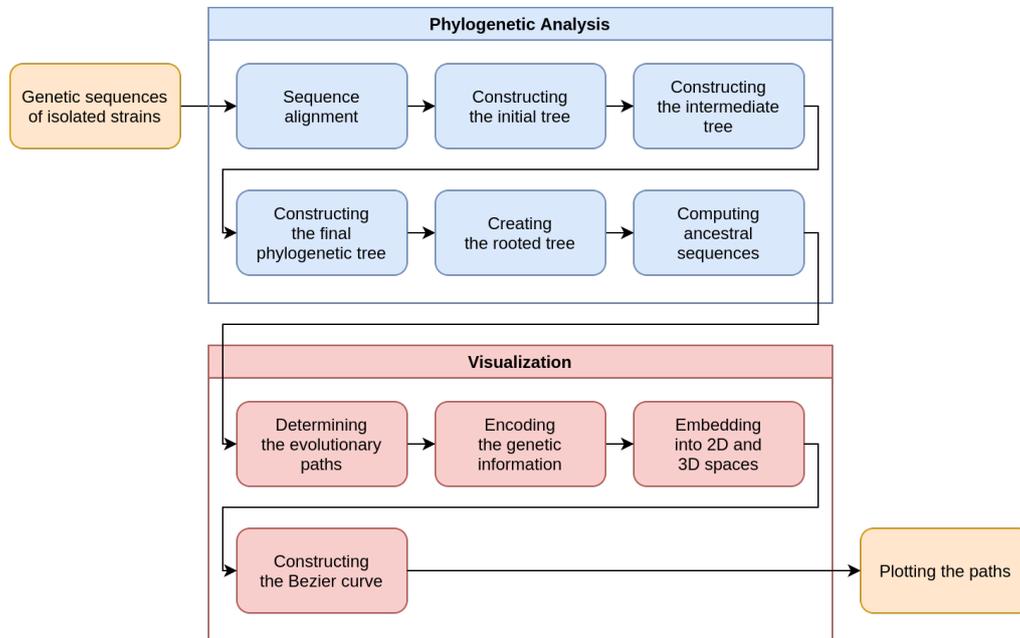


Figure 1: Overall schema of the proposed approach. It contains two general steps: phylogenetic analysis; and constructing the visualization.

Before performing any computations, it is necessary to conduct an alignment procedure for selecting a fragment of the genetic sequence with maximum coverage of the information. Alignment is an inseparable stage for constructing the phylogenetic tree in our task. Several methods perform the alignment, among which Multiple Alignment using Fast Fourier Transform (MAFFT) [18] is advantageous for large data sets. The aligned sequences file is further given to the Randomized Axelerated Maximum Likelihood (RAxML) [19] program for generating the phylogenetic tree. It is known that stochastic models (such as maximum-likelihood) are more desirable for biological research, but they often suffer from low computational efficiency. The main advantage of RAxML is its speed in parallel computation of the best maximum-likelihood score; this makes it a suitable choice for working with large-scale data sets.

RAxML gives a wide choice of models, for both nucleotide and amino acid sequences, used to generate the tree. Computing the ancestral sequences by RAxML requires a rooted tree. Therefore, the next step is making a rooted from an unrooted tree by setting the flag '-f I' in RAxML and constructing the ancestral sequences by setting the flag '-f A'.

An evolutionary path starts from the root, passes to the internal nodes (ancestors), and ends in an isolated strain (i.e., leaf of tree). Thus, the total amount of paths is equal to the number of leaves in the tree. The first step in our visualization is encoding the genetic information of each

path. Since we use nucleotide sequences, the alphabet of which contains four nucleotides (A, C, G, T) and gap (-), we apply the one-hot-encoding in Table 1 to represent the information in the numerical domain:

Table 1

One-hot-encoding for converting the genetic information of nucleotide sequence into a binary sequence.

Letter	Code
A	(1,0,0,0)
C	(0,1,0,0)
G	(0,0,1,0)
T	(0,0,0,1)
'-' Gap	(0,0,0,0)

Some positions in the sequence are conserved and non-informative, so we remove them from the further computation. Finally, we obtain a binary matrix, whose rows indicate the nodes in the path; its columns are the encoded genetic information. We apply the embedding method to the matrix to project paths from multi-dimensional space into 2D or 3D spaces. Our preliminary results indicated that the multi-dimensional scaling outperforms others among several methods of visualization. The achieved 2D path is represented in the form of the Bezier curve in a 2D plot. In the next section, we apply our computational pipeline (presented in Figure 1) to visualize the evolutionary paths of the influenza virus.

3. Experiment Setup & Results

3.1. Data Preparation

We downloaded more than 90,000 nucleotide sequences of influenza virus subtype H3N2, isolated from 1967-2021, from the GISAID database (Global Initiative on Sharing All Influenza Data) [20]. After filtering out duplicate entries, aligning sequences, cleaning the database, and removing sequences with ambiguous nucleotides, we obtained more than 30,000 strains. The strains further were sorted by their isolation year. A sample of up to 200 entries was selected for each year. The final data set was created by gathering all samples, and it included about 5,000 strains isolated in the period from 1968-2021. Note that some earlier years have less than 200 samples after data preprocessing.

3.2. Constructing The Phylogenetic Tree

Maximum-likelihood tree construction consists of three sequential procedures: generating initial, intermediate, and final trees. The initial maximum-likelihood tree was generated using Fasttree [21] with the Juke-Cantor model. In contrast, RAxML was applied to create the intermediate tree from the initial one with a generalized time-reversible (GTR) model and the rapid hill-climbing mode. The obtained tree was evaluated under the GAMMA model of rate heterogeneity modeling. Next, the final, refined maximum-likelihood tree was generated by RAxML from the intermediate tree under GTR and GAMMA models.

3.3. Reconstructing Ancestral Sequences

In order to compute ancestral sequences by RAxML, the tree must be rooted. Since the final tree was not rooted, we apply RAxML with the flag '-f I' to generate the rooted version. The child-parent relationship between tree nodes was extracted using 'Phylo' modules from the Biopython package[22].

3.4. Visualization

The child-parent relationship allows us to create the evolutionary path, which starts from the root and ends in a tree leaf. We encode the genetic information of each path by applying one-hot-encoding presented in the Methodology. We removed conserved sites, as they are not informative. We embedded the multi-dimensional (encoded) representation of strains into 2D space, to visualize paths, by employing MDS.

Given n point $X = \{x_1, x_2, \dots, x_n\}$ in a high dimensional space (p dimensions) and their distance affinity matrix D , MDS aims to find point $Y = \{y_1, y_2, \dots, y_n\}$ in a space of lower dimension (q dimensions) such that:

$$\min_Y \sum_{i=1}^n \sum_{j=1}^n (d_{i,j} - \hat{d}_{i,j})^2$$

Note that \hat{D} is the distance affinity matrix for points in lower dimensional space (containing $\hat{d}_{i,j}$ the distance between y_i, y_j).

Suppose an evolutionary path P includes n strains of length L and is presented by vector:

$$P_{raw} = [X_1, X_2, \dots, X_n]$$

where $X_i, i \in \{1, 2, \dots, n\}$ represents the sequence of nucleotides as follows:

$$X_i = [x_{i,1}, x_{i,2}, \dots, x_{i,L}]$$

where

$$x_{i,j} \in \{A, C, G, T, -\}, \quad j \in \{1, 2, \dots, L\}$$

Since conserved amino acid positions are uninformative, they are removed from the sequence. Suppose l is the number of positions, each of which has at least one mutation. By applying the one-hot-encoding from Table 1 to each strain, we obtain the encoded path:

$$x_{i,j} \xrightarrow{\text{Encoding}} [y_{i,e}, y_{i,e+1}, y_{i,e+2}, y_{i,e+3}]$$

$$P_{encoded} = [Y_1, Y_2, \dots, Y_n]$$

where

$$Y_i = [y_{i,1}, y_{i,2}, \dots, y_{i,k}]$$

and

$$y_{i,j} \in \{0, 1\}, \quad i \in \{1, 2, \dots, n\}, \quad j \in \{1, 2, \dots, k\}, \quad k = l \times 4$$

Each tree node (internal or terminal) can be represented as a point in high dimensional space ($l \times 4$ dimensions). We map the data points of the path from high dimensional space into 2D or 3D space by employing MDS:

$$P_{embedded} = [Z_1, Z_2, \dots, Z_n]$$

where Z_i is either a 2D or 3D point. The final coordinates were further smoothed to form a Bezier curve and visualized in a 2D plot. A typical rendering of evolutionary paths is presented in Figure 2. Note that each path is individually embedded into the lower space. Thus the visualization of a path is independent from others.

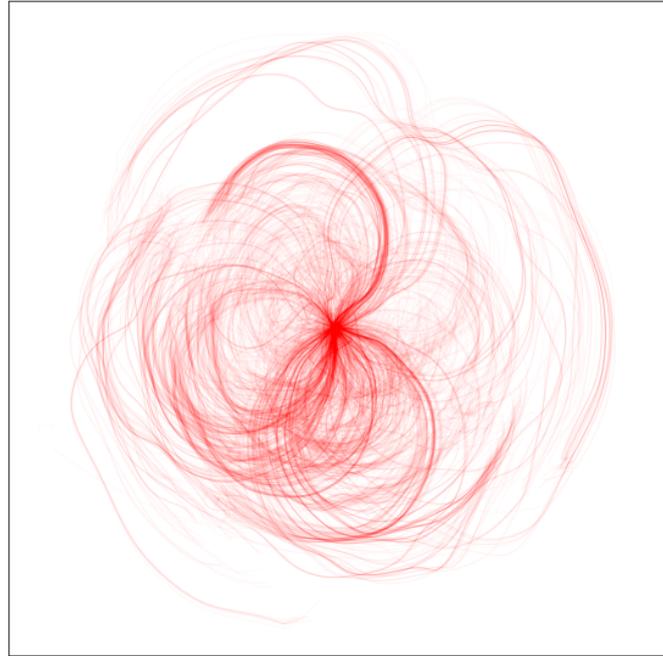


Figure 2: Visualization of evolutionary paths. Note that all paths start from the root, located at the center of the plot.

In order to evaluate how 'genetically-close' and 'genetically-far' strains are characterized in the new space, we visualized the paths of the closest and most distant strains of a randomly selected year (2016) in Figure 3. Note that the closest strains have only a one-mutation difference. This high degree of genetic similarity leads to almost the same curve and a slight change of their positions. In contrast, from the right plot of Figure 3, as expected, we see that the more differences between the genetic sequences, the more distance between their paths.

We applied four different algorithms to visualize the resultant Bezier curves of the sample strains. The algorithms include multi-dimensional scaling, t-SNE, Isomap [23], and kernel PCA [24]. A randomly selected sample of 200 strains was visualized by the aforementioned embedding algorithms. We applied different values for hyperparameters of t-SNE method. In our preliminary results, MDS outperformed others by providing a more clear, and less crowded,

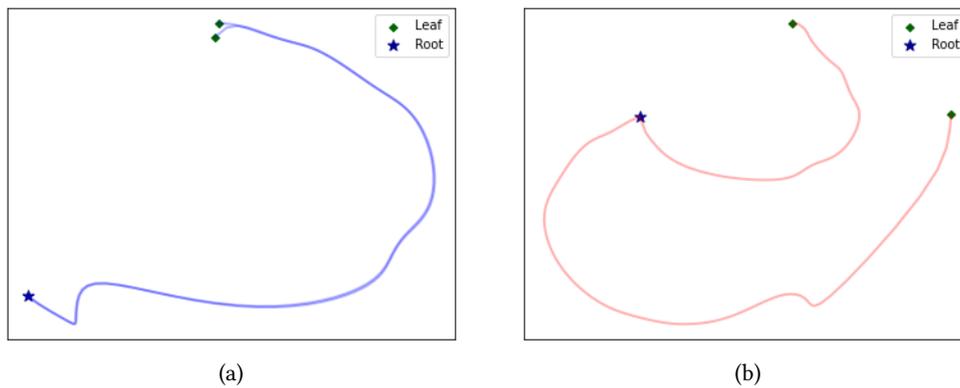


Figure 3: Visualization of evolutionary paths for (a) genetically-close and (b) genetically-far strains. As expected, the plots show that close strains have a more similar path than far strains.

visualization. We believe a suitable choice of t-SNE hyperparameters may provide a better visualization, which is the subject of our future work. It is worth mentioning that sometimes the visualization includes outlier strains, which can be due to: low sequence quality; single events that cause deleterious (for the virus) genetic variation; or false information about isolation date. Although this happened rarely, we removed the paths of such strains from the visualization.

Visualization of evolutionary paths can be used to assess the variation among a set of sequences. To demonstrate the power of such visualization, we plotted the paths of randomly selected strains for the years 2001, 2008, and 2019. Figure 4 indicates that there are more differences between the strains, and consequently their paths, in 2008. In contrast, the sample strains in 2001 and 2019 feature more similar paths, forming clusters in the visualization.

To increase visualization quality, we provide an interactive 3D representation equipped with virtual reality. It is built using *Viewzavr*, a framework for constructing visualizations. It connects three levels of programming: a language level; visual programming; and an end-user interaction (which is also considered programming). A randomly selected sample, of up to 100 paths from each year, are plotted in Figure 5. The isolated strains (leaf of the phylogenetic tree) are presented with orange spheres. This visualization is available online at github.com/viewzavr/vr-flu-galaxy.

Generally speaking, our results indicate that a comparison of two strains can be enhanced by incorporating information on their ancestors. Thereby, such comparison does not individually consider two strains, but it evaluates them in a chain of events, where each event represents an ancestor. Indeed, the hierarchical differences between two strains can be reflected through their paths. We believe that incorporating the evolutionary history of strains may provide a better characterization and improve the quality of viral evolution modeling.

4. Conclusion

Analysis of viral evolution is essential in the fight against viruses. This paper proposes a visualization method for viral evolution based on reconstructing the phylogenetic tree and ancestral sequences. Our method projects an evolutionary path of the phylogenetic tree into

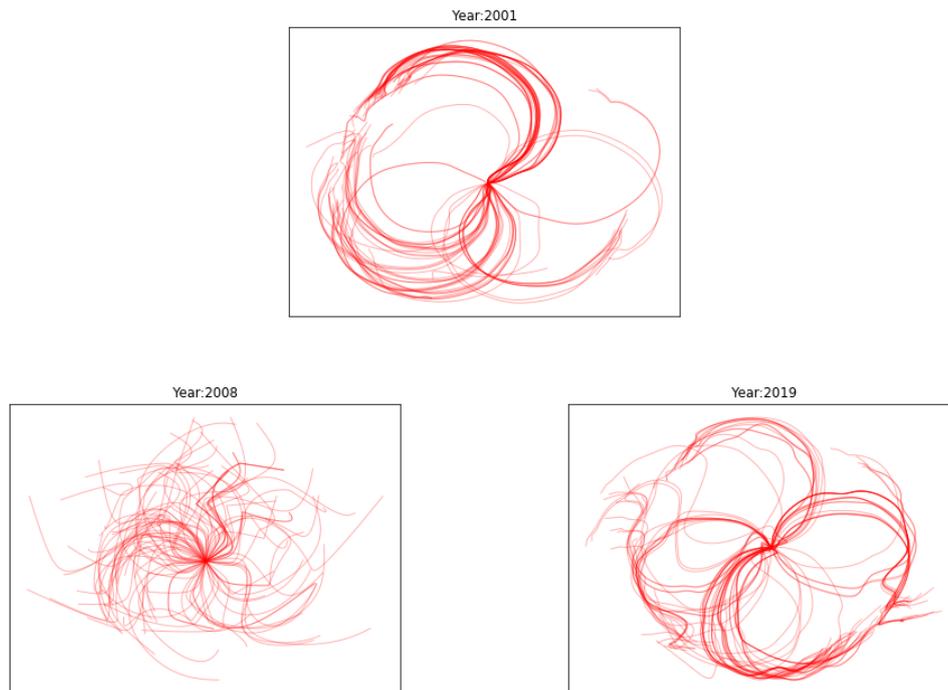


Figure 4: Visualization of evolutionary paths for a randomly selected sample of 100 strains. Both 2001 and 2019 samples indicate more similarity between strains, while 2008 samples express more differences between paths.

a 2D or 3D space by incorporating the genetic information of nodes located in the path. The suggested method can serve as an exploratory tool to visually survey viral variation. The hierarchical representation of a strain provides additional information, which may improve the characterization of strains through their paths. The (euclidean) distance, between the reference and test viruses in the low dimensional space, can be beneficial in modeling antigenic evolution. From a technical perspective, the only considerable drawback to our approach is the computational complexity of constructing a phylogenetic tree from a large number of viruses and projecting its nodes from the high dimensional into the low dimensional space.

Although the paper's case study is the influenza virus hemagglutinin protein, and our result represents a partial visualization of its evolution, we plan to perform a more comprehensive visualization using the entire genome. Note that the approach is easily extendable to other viruses. In addition, alternative representations can be created by applying our method to amino acid sequences, instead of nucleotides, and performing analysis with simplified amino acid alphabets. This allows us to visualize and study evolution from various viewpoints, such as hydrophobicity. Future work needs to be done to set appropriate criteria (metrics, limits, etc.) that permit automatic assessment and recognition of regular versus irregular viral visualization patterns.

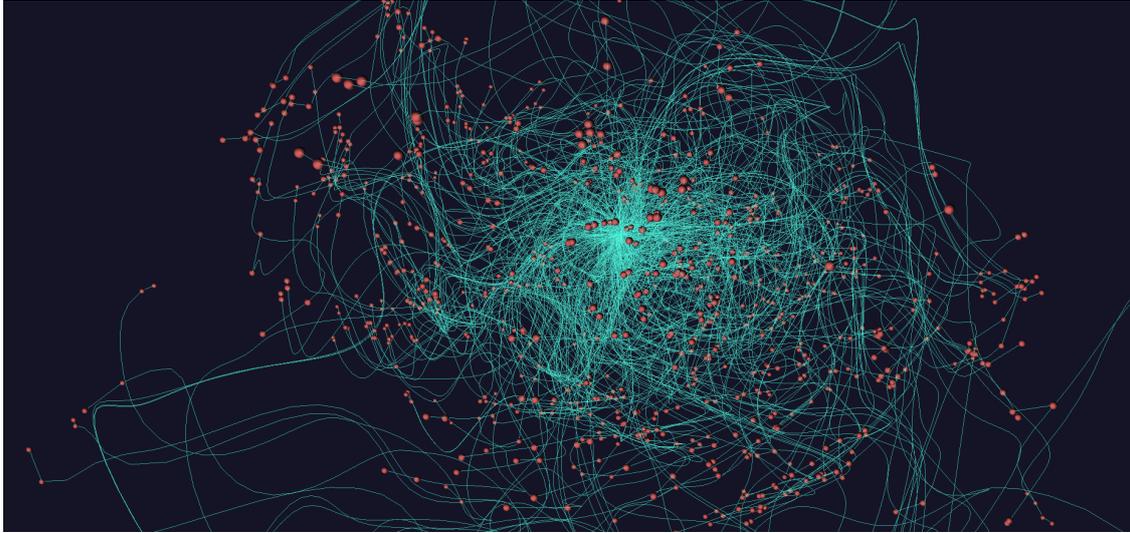


Figure 5: An interactive, 3D visualization of evolutionary paths. The visualization was generated by randomly selecting up to 100 strains from each year (1968-2020). Each isolated strain is indicated at the end of the path by an orange sphere.

Acknowledgments

The reported study was funded by the Russian Foundation for Basic Research, project number 19-31-60025.

Our work was performed using the “Uran” supercomputer (IMM UB RAS).

References

- [1] W. H. Organization, et al., Global influenza strategy 2019-2030 (2019).
- [2] M. Forghani, M. Khachay, Convolutional neural network based approach to in silico non-anticipating prediction of antigenic distance for influenza virus, *Viruses* 12 (2020) 1019. doi:10.3390/v12091019.
- [3] T.-M. Rhyne, Does the difference between information and scientific visualization really matter?, *IEEE Computer Graphics and Applications* 23 (2003) 6–8. doi:10.1109/MCG.2003.1198256.
- [4] G. E. Jordan, W. H. Piel, Phylowidget: web-based visualizations for the tree of life, *Bioinformatics* 24 (2008) 1641–1642. doi:10.1093/bioinformatics/btn235.
- [5] M. Forghani, P. Vasev, V. Averbukh, I. Ras, Three-dimensional visualization for phylogenetic tree, *Scientific Visualization* 9 (2017) 59–66. doi:10.26583/sv.9.4.06.
- [6] M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, *Journal of molecular evolution* 16 (1980) 111–120. doi:10.1007/bf01731581.
- [7] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing

- phylogenetic trees., *Molecular biology and evolution* 4 (1987) 406–425. doi:10.1093/oxfordjournals.molbev.a040454.
- [8] W. M. Fitch, E. Margoliash, Construction of phylogenetic trees, *Science* 155 (1967) 279–284. doi:10.1126/science.155.3760.279.
- [9] E. Mayr, Cladistic analysis or cladistic classification?, *Journal of Zoological Systematics and Evolutionary Research* 12 (1974) 94–128. doi:10.1111/j.1439-0469.1974.tb00160.x.
- [10] A. Soares, R. Râbelo, A. Delbem, Optimization based on phylogram analysis, *Expert Systems with Applications* 78 (2017) 32–50. doi:10.1016/j.eswa.2017.02.012.
- [11] J. Podani, The coral of life, *Evolutionary Biology* 46 (2019) 123–144. doi:10.1007/s11692-019-09474-w.
- [12] W. T. Harvey, D. J. Benton, V. Gregory, J. P. Hall, R. S. Daniels, T. Bedford, D. T. Haydon, A. J. Hay, J. W. McCauley, R. Reeve, Identification of low-and high-impact hemagglutinin amino acid substitutions that drive antigenic drift of influenza a (h1n1) viruses, *PLoS pathogens* 12 (2016) e1005526. doi:10.1371/journal.ppat.1005526.
- [13] R. A. Neher, T. Bedford, R. S. Daniels, C. A. Russell, B. I. Shraiman, Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses, *Proceedings of the National Academy of Sciences* 113 (2016) E1701–E1709. doi:10.1073/pnas.1525578113.
- [14] K. Ito, M. Igarashi, Y. Miyazaki, T. Murakami, S. Iida, H. Kida, A. Takada, Gnarled-trunk evolutionary model of influenza a virus hemagglutinin, *PloS one* 6 (2011) e25953. doi:10.1371/journal.pone.0025953.
- [15] M. A. Cox, T. F. Cox, Multidimensional scaling, in: *Handbook of data visualization*, Springer, 2008, pp. 315–347. doi:10.1007/978-3-540-33037-0_14.
- [16] C. A. Steinparz, A. P. Hinterreiter, H. Stitz, M. Streit, Visualization of rubik’s cube solution algorithms., in: *EuroVA@ EuroVis, 2019*, pp. 19–23. doi:10.2312/eurova.20191119.
- [17] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., *Journal of machine learning research* 9 (2008).
- [18] K. Katoh, D. M. Standley, Mafft multiple sequence alignment software version 7: improvements in performance and usability, *Molecular biology and evolution* 30 (2013) 772–780. doi:10.1093/molbev/mst010.
- [19] A. Stamatakis, Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, *Bioinformatics* 30 (2014) 1312–1313. doi:10.1093/bioinformatics/btu033.
- [20] Y. Shu, J. McCauley, Gisaid: Global initiative on sharing all influenza data—from vision to reality, *Eurosurveillance* 22 (2017) 30494. doi:10.2807/1560-7917.ES.2017.22.13.30494.
- [21] M. N. Price, P. S. Dehal, A. P. Arkin, Fasttree 2—approximately maximum-likelihood trees for large alignments, *PloS one* 5 (2010) e9490. doi:10.1371/journal.pone.0009490.
- [22] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, et al., Biopython: freely available python tools for computational molecular biology and bioinformatics, *Bioinformatics* 25 (2009) 1422–1423. doi:10.1093/bioinformatics/btp163.
- [23] M. Balasubramanian, E. L. Schwartz, J. B. Tenenbaum, V. de Silva, J. C. Langford, The isomap algorithm and topological stability, *Science* 295 (2002) 7–7. doi:10.1126/science.

295.5552.7a.

- [24] S. Mika, B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz, G. Rätsch, Kernel pca and de-noising in feature spaces., in: NIPS, volume 11, 1998, pp. 536–542.

A. Online Resources

The 3D visualization of influenza viruses are available via github.com/viewzavr/vr-flu-galaxy