2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS) | 978-1-6654-0170-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICHMS53169.2021.9582636

An Explorative Tool for Mutation Tracking in the Spike Glycoprotein of SARS-CoV-2

Johannes Schwerdt
Department of Technical and Business Information Systems
Otto von Guericke University Magdeburg
Magdeburg, Germany
johannes.schwerdt@ovgu.de

Aljoscha Tersteegen & Pauline Marquardt
Institute of Medical Microbiology and Hospital Hygiene
Medical Faculty of the Otto-von-Guericke University
Magdeburg, Germany
GivenName.Surname@med.ovgu.de

Achim J. Kaasch
Institute of Medical Microbiology and Hospital Hygiene
Medical Faculty of the Otto-von-Guericke University
Magdeburg, Germany
achim.kaasch@hhu.de

Andreas Nürnberger
Department of Technical and Business Information Systems
Otto von Guericke University Magdeburg
Magdeburg, Germany
andreas.nuernberger@ovgu.de

***Abstract*—Interactive Information Visualization and Human Computer Interaction provides useful support for inexperienced user and experts as well. On one hand visualization provides an informative overview of data and on the other hand interaction encourages users for exploration within it. We chose the challenge of a specialized/expert scenario to build a pipeline that provides interactive visualizations to encourage users for further exploration. We chose the complex task of a phylogenetic analysis of SARS-CoV-2 genomes for mutation tracking. The proposed pipeline hides the mathematical details while providing complex information visually and intuitively. In our proof-of-concept we analyzed four variants of concern and identified mutations in the spike glycoprotein with more than 70% precision and 77% recall in reference to the reports of the Centers for Disease Control and Prevention.**

***Index Terms*—SARS-COV-2, Spike Glycoprotein, Mutation Detection, Interactive Information Visualization**

I. Introduction

The current COVID-19 pandemic reminded the entire scientific community to work together in a combined effort. Several scientific fields work on a shared focus but often face problems while exchanging ideas, models and analysis tools. Practitioners in the COVID-19 research are associated to biology, chemistry or clinical research. With the increasing magnitude of publicly available data, these practitioners are forced to deal with problems of the computer science community and complex analysis tools from the bioinformatics community. A fundamental aim of this research comprises the task to identify mutations within the SARS-CoV-2 genome and track variants of shared mutation profiles. The detection of mutations in an evolving genome is a challenging task on its own and tracking individual variants within a population is an even harder one. Such an analysis comprises several complex steps, such as data acquisition of genomic sequences, data preprocessing and data analysis, e.g. hierarchical clustering, to characterize mutations in its variants. During the task experts gain overview of data, their intrinsic similarities and eventually draw meaningful conclusion. Such a process is abstractly described in Fig. 1. While experts have years of practice and expertise, less routine practitioners face high barriers towards understanding and solving this tasks. We propose a pipeline comprising visual support on several levels to encourage users exploration and lower such barriers. This work presents an end-to-end solution for the analysis while providing interactive visualization to guide individual searches of a user. For sure, this paper is not intended for experts in the field but for the new practitioners that the current times motivated for the task. Nonetheless, we are confident enough to state that experts might be interested in at least aspects of the proposed pipeline and its explorative possibilities.

In summary, the contribution of this paper includes:

- Explorative visualization for mutation tracking using genomic data
- Providing explainable solutions to encourage users interaction and exploration
- Detailed evaluation of the proposed method in a proof-of-concept scenario

The structure of the paper is as follows. We start the paper with a small description of the related work in section II. In section III we explain the algorithmic details of our pipeline, which comprises data acquisition & preprocessing, hierarchical clustering and cluster extraction & interpretation. In section IV we propose and apply our visualization scheme to state our findings and results. Finally, we end the paper with our conclusion in section V.

II. Related Work

Most of the related work are the inspiration and/or part of the approach proposed later in this work. A prominent tool providing support and solutions for the same task is

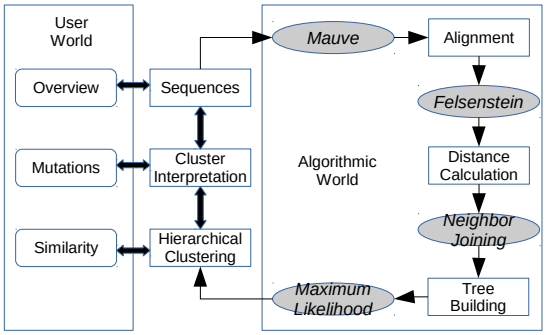


Fig. 1: The presented pipeline follows the aim of mutation analysis in biosequences. User's have specific expectations while working on this task (left), while machines provide input/output [i/o] (arrow direction) (right). An individuals cognition is able to connect the i/o and draw conclusions about the problem (double arrow).

Nextstrain/Nextclade [1]. *Nextclade* is a full end-to-end solution for mutation calling, phylogenetic analysis and assigning individual sequences into groups, called clades. With an easy to use web solution, one can upload SARS-CoV-2 genomes for a complex analysis. Users are provided with information about sequence quality scores and individual mutations. The sequences are put into a phylogenetic tree and visually highlighted with the aim that the user can inspect genome similarities to previously analyzed sequences. Additional graphics present position-wise sequence diversities, providing the user with information about mutation rich positions within the genome. The provided visualizations capture crucial information about the data and their intrinsic similarities. But the visual support does not scale well with a large amount of data and a user is limited in his/her interaction with it. This discourages users for further exploration within the presented information. Several phylogenetic tree viewers exist, such as *Interactive Tree Of Life (iTOL)* [2], *Dendroscope* [3], *A Tree Viewer* [4] and its successor *Archaeopteryx*. While iTOL being one of the best visualization tools for that purpose, it does not scale too well with large data sets. Options for interaction and exploration within the data are limited. On the other hand, network visualization toolboxes outside the community seemed to be promising candidates for alternatives. *visNetwork* is a javascript library that can easily be adapted to individual needs. We adapted it for the representation of phylogenetic trees and used functionalities to encourage exploration within the tree. The visualization is integrated into a prototypical analysis pipeline comparable to *Nextclade*. We hope our alternative pipeline gives inspiration for new development and complements existing solutions with new perspectives.

III. Methods

In our analysis we acquired SARS-CoV-2 genomes, preprocessed them in a comparable format and clustered them hierarchically. This section focuses on the aspects we called *Algorithmic World* in Fig. 1

A. Data & Preprocessing

The *GISAID Initiative* [5] provides access to download the sequenced SARS-CoV-2 genomes extracted from infected individuals. Because of the sheer magnitude of the data and the associated effort in calculation, we randomly sub-sampled the downloaded data to the arbitrary size of 3,506 genomes. The virus mutates over time manifesting changes in its genome in form of substitutions, insertions and deletions. A substitution defines a change of a (ribo)nucleotide towards another, while insertions adds and deletions removes (ribo)nucleotides. For that reason genomes do not share unified shifts without adequate preprocessing. The problem statement of a multiple sequence alignment (MSA) describes the challenge to identify such an unit shift. To generate sequences with the same length while maximizing matching/conserved regions gaps will be inserted. The result is called an alignment and sequences are put into their unified shift. The bioinformatics community provides a plethora of models, tools and software for that task with individual strengths and benefits. We chose *Mauve* [6] because it is a multiple genome alignment tool well suited for research of comparative genomics and the study of genome-wide evolutionary dynamics.

B. Hierarchical Agglomerative Clustering

To gain insights of the mutation dynamics within the SARS-CoV-2 genomes, we can apply a hierarchical clustering on these sequences that have an unified shift in their alignment after preprocessing. Therefore, we can implement pairwise distance functions, like the hamming distance, to create a distance matrix for a subsequent clustering. Such naive distance functions are unable to reconsider back-and-forth mutations ('double' mutations that cancel out). Meaningful distance functions to model DNA/RNA evolution are the foundation of molecular phylogenetics. Prominent models are *JC69* [7], *K80* [8], *F81* [9] and *HKY85* [10]. Adequate selection of these models can be done by statistical testing. Hierarchical clustering is a prominent tool in the data science community to extract knowledge from data. Nonetheless, the hierarchy we aim to identify follows an evolutionary interpretation. For that reason, we choose specialized algorithms from the bioinformatics community, e.g. *Neighbor-Joining* [11] and *Maximum-Likelihood* [9] for hierarchical clustering. Such clustering can be interpreted as phylogenetic trees representing evolutionary processes connecting the individual biosequences. *phangorn* [12] is a toolbox written in R for exactly that scope. We used *phangorn*'s functionalities on the MSA to derive a corrected distance matrix with F81 and build the hierarchical clustering with Neighbor-Joining. Subsequently, we refined the clustering with a combination of F81 and Maximum-Likelihood, similar to Nguyen et al. [13].

C. Cluster Extraction & Interpretation

To derive meaningful semantics from the clustering, we further enhance our data with meta information. The input data comprises SARS-CoV-2 genomes that can be assigned to mutation families. Two prominent tools are used to derive such assignments, *Nextclade* [1] (clade assignments e.g. *20I/501Y.V1*) and *Pangolin* [14] (lineage assignment e.g. *B.1.1.7*). This meta information will be assigned to the leaf nodes in the clustering to be used later in the graph visualization. All inner nodes (mergings in the clustering) also need interpretable semantics assigned to them. For that reason we slightly reverse-engineered aspects of the tree building process. A central part of the tree building is the Maximum-Likelihood method according Felsenstein, see Algo. 1. Given a tree structure (nodes and edge-lengths) and a model of evolution (e.g. F81), likelihoods can be calculated accordingly. Because of the recursive nature of the algorithm each inner node comprises its own likelihood that can be used to create a prototypical 'cluster-genome'. The maximized likelihood reflects a sequence best representing all its child nodes. This sequences can be seen as an evolutionary ancestor of the nodes children or simply its cluster representation. Therefor creating meaningful semantics for cluster interpretation.

Algorithm 1 Felsenstein's algorithm for likelihood (extracted with adaptions from [15]).

procedure LIKELIHOODFORPOSITIONU
  Initialisation:
    N sequences at their u-th position $\{x_u^1, ..x_u^N\}$
    Tree T with nodes k and edge-lengths t_k
    Model of evolution $P(.|., t_i)$ and $P(.)$
  Recursion:
  Compute likelihood $P(L_k|a)$ for nucleotide a:
  if k is leaf node **then**
    $P(L_k|a) = \delta(a = x_u^k)$
  if k is inner node **then**
    $P(L_k|a) = \sum_{b,c} P(b|a, t_i) \cdot P(L_i|b)$
    $\cdot P(c|a, t_j) \cdot P(L_j|c)$
  Termination:
  $P(x_u|T, t.) = \sum_a P(L_{root(T)}|a) \cdot P(a)$

IV. Results

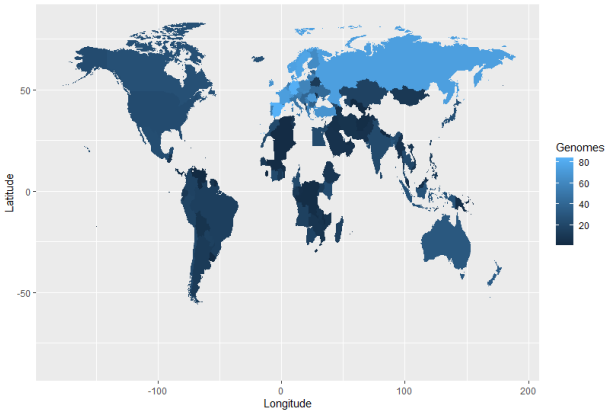
This section is structured in the following. Sub-section IV-A describes the visualization of accessible meta information to create an overview of the data. Sub-section IV-B visually analyzes the hierarchical clustering process itself. We describe how to draw conclusion from it and how to identify obscurities. Sub-section IV-C describes the analysis of clusters for mutations in a visual way. This section focuses on the aspects we called *User World* in Fig. 1

A. Visualization of the Overview

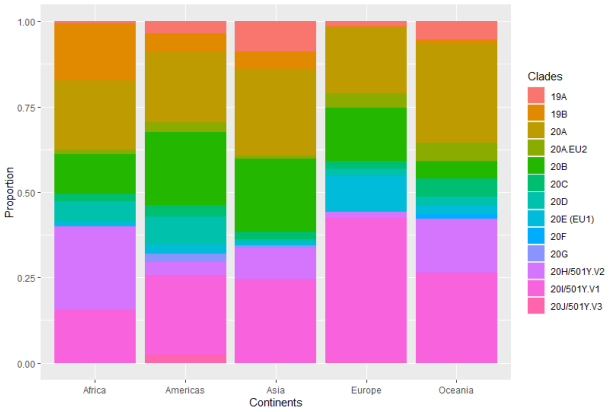
The first step of this pipeline is the presentation of an overview using accessible meta information. A user needs as much knowledge as possible about the data at hand. The individual information need of users should be satisfied as soon as possible. Meta information of data might be a useful filter for undesired input or to precisely focus towards a subsample of data. For an easy and accessible overview we provide the user with the following information. The SARS-CoV-2 genomes are stored in the *FASTA* format comprising information of the extraction location in the headers. Further on, tools like *Nextclade* proved assignments characterizing genomes in groups (so-called clades). Such sources of meta information can easily be combined and visualized to create a comprehensive overview, see Fig. 2. Confronted with this information the user might redefine some aspects of subsequent analysis. We can easily see that Europe and Russia provided a magnitude of sequences. This does not indicate a higher level of infectious activity but simply more provided genomes. Clade proportions across the continents are different but comparable. While looking at individual countries, we can observe more fluctuation. The graphics can easily be produced by ca. 20 lines of code in the language R using the libraries *ggplot2*, *maps* and *countryCode*. They provide a comprehensive overview of groupings in geolocations. Such visualizations increase in value when combined with date information to observe changing dynamics.

B. Visualization of the Similarity

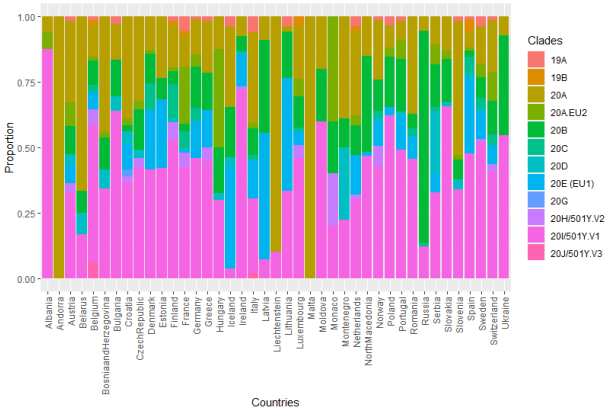
The next step of this pipeline is the presentation of sequence similarities to identify clusters of comparable mutations within genomes. For that reason, users should be able to fully explore and interact with the space of presented information. The human intuition might even be capable to uncover clusters or outliers that have algorithmically not been treated adequately. With such visualization, the user might redefine aspects of the previous analysis or focuses his/her subsequent analysis on particular sub-graphs within the complete clustering. We chose *visNetwork* because we think it encourages graph exploration by providing user-friendly interactions such as navigation, zooming, clicking and several node or attribution selections. The functionalities of the framework could be easily adapted to the task at hand because of the frameworks superb documentation. Data exchange between the visualization and the clustering tools could be done with a few lines of code. *phangorn*'s data structures provide all functions needed to extract node and edge information for the transformation into *visNetwork*. We focus our analysis on the *spike glycoprotein* region. Mutations in this region have been reported to affect the virus fitness [16]. The *National Center for Biotechnology Information* provides information of genes within the genome and we cut the alignment in reference to the original SARS-CoV-2 genome. The hierarchical clustering of the spike glycoprotein region can be seen in Fig. 3. Each leaf node is an individual SARS-CoV-2 sequence and connecting edges encode sequence similarities by their length. The closer nodes are connected within the network, the higher is their sequence similarity. To incorporate more information into this visualization, we color encoded the leaf nodes with their clade



(a) Provided genomes by country



(b) Proportion of clades across continents



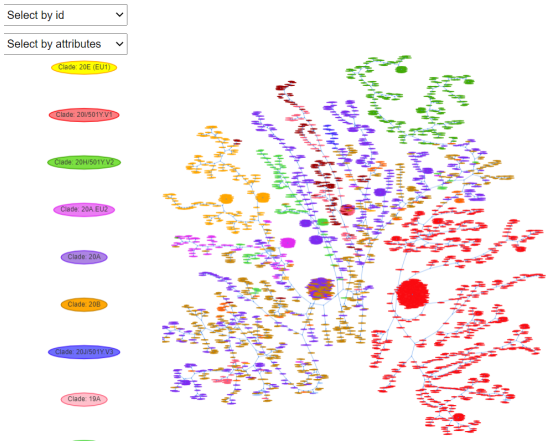
(c) Proportion of clades across European countries

Fig. 2: Geolocations of provided SARS-CoV-2 genomes. Genomes were assigned to clades via *Nextclade* and can be compared across several locations.

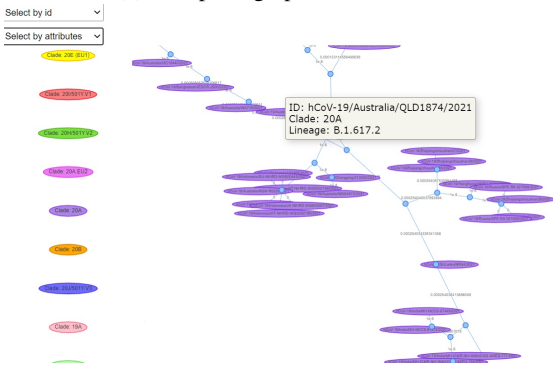
assignment from *Nextclade*. Lineage assignments by *Pangolin* are added as attributes and can be highlighted using a sidebar. The same can be done with node IDs. By clicking on the nodes the user is provided with a summary of information about the data geolocation, extraction date, clade and lineage assignment. Such an interaction rich tool encourages users to explore the complete clustering and interact with it. In the domain of medical investigation such interactive visualization could provide overview and context. Especially, when clinical factors are added such as (the genome was extracted from a) patient with certain age, sex and/or medical record. By visual inspection of the graph, we see that the clustering captures sub-graphs marked with the same node color encoding as provided by *Nextclade*. By highlighting individual lineages provided by *Pangolin*, we observe again connected sub-graphs clusters. This indicates that our clustering captures in essence the same information and provides us with confidence in the proposed pipeline. We can further see that the clustering comprises 'big bubbles' of sub-clusters. These bubbles might be stable mutation spots that provide an equilibrium state for the virus variants. An alternative explanation could be an artefact, caused by overrepresentation due to biased sequencing. Several paths branch away from these spots as one could image when thinking of evolutionary events. A detailed analysis of these branches vs. bubbles might be a fruitful future research question to gain deeper understanding of the mutation dynamics within these clusters.

C. Visualization of the Mutations

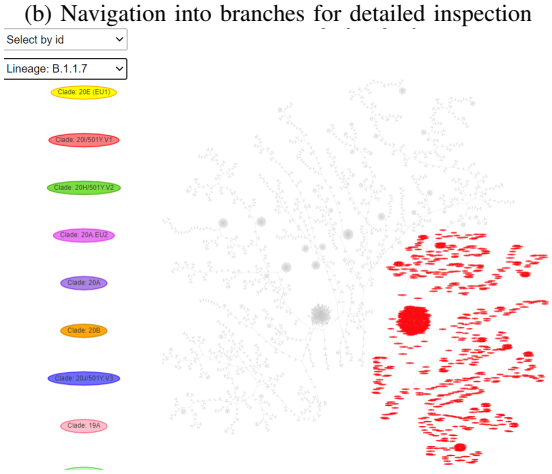
The last step of this pipeline enriches the clustering with a semantic interpretation. Users should be provided with a clear solution so the individual can decide on his/her own if the information need is fulfilled or not. If the results are not satisfactory, the user can redefine aspects of the previous analysis, such as data acquisition, filtering or preprocessing. If the information need is indeed satisfied, the analysis and exploration is finalized. In case when we have no clear semantic interpretation of the clusters, in all honesty, we simply have a graphic or algorithm of no practical use. For that reason, we propose a semantic interpretation, use it for predictions and evaluate it in a proof-of-concept scenario. We focus the analysis on the currently reported variants of concern, namely *B.1.1.7*, *B.1.351*, *B.1.617.2* and *P.1*. These variants are well researched and a set of their mutations is publicly available. This knowledge provides us with the possibility to evaluate our pipeline later on. Current research focuses on mutations on the protein sequence rather than the nucleotide sequence. Therefore, we translate the genes into their protein sequences when doing the comparisons. As already described in subsection IV-B, we are able to highlight sub-graphs of specific lineages and recognized sub-trees of close proximity forming dense clusters. The toolbox provides easy navigation to identify the root-node and we extract the SARS-CoV-2 sequences under this node. Additionally, we inferred the prototypical 'cluster-genome' in that root-node by the Maximum Likelihood version of Algo. 1. For an easily understandable visualization of these clustered sequences, we used *AliView* [17]. This toolbox enables the user to highlight the differences towards a reference sequence, e.g. the original SARS-CoV-2 version. Fig. 4 shows a short snippet of the results. It indicates that the clustering was able to identify highly similar sequences and the inference of the prototypical 'cluster-genome' represents their cluster adequately.



(a) Complete graph visualization



(b) Navigation into branches for detailed inspection



(c) Highlight sub-graphs with node attributes (e.g. lineages)

Fig. 3: Visualization of the hierarchical clustering within the *spike glycoprotein* region. Leaf nodes represent individual genes and inner nodes represent mergings during the clustering. The color of nodes encodes the clade assignment from *Nextclade* while lineage assignments from *Pangolin* can be selected as attributes.

We compare these 'cluster-genome's with the original SARS-CoV-2 reference to identify mutations. Mutations like *N501Y* are defined as followed: at position 501 the original strain has a 'N' amino acid while the mutant has a 'Y'. The

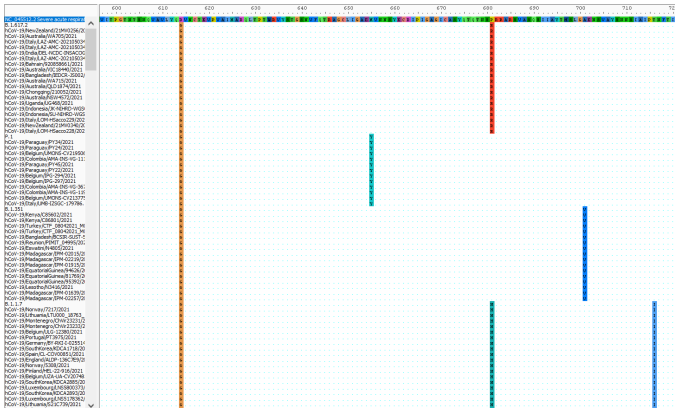


Fig. 4: Alignment of the *spike glycoprotein*. Differences to the wild-type SARS-CoV-2 sequence are highlighted. Entries named like variants of concern have been calculated by the Maximum-Likelihood version of Algo. 1 as 'cluster-genome's.

Centers for Disease Control and Prevention (*CDC*) [18] and the *Robert Koch-Institut* (*RKI*) [19] report variants of concern with known mutations. Identified mutations by the pipeline can easily be compared with them. Our results can be found in Tab. I. The *CDC* reports 10 mutations for *B.1.617.2*. The proposed approach agreed in 9 but not on *R158G* because it was identified as deletion *158del*. In that reference, the pipeline achieved 90% in precision and recall. For *P.1.* the *CDC* lists 11 mutations and the algorithm agreed in 10 but missed *D138Y*. In addition *V1176F* was identified, such as listed by the *RKI*. Pessimistically, we treat it as 90.9% in precision and recall. *B.1.351* has 10 reported mutations. The pipeline identified 7 of them but missed all deletions: *241del*, *242del* and *243del*. This results in 70% precision and 100% recall. *B.1.1.7* comprises 13 reported mutations while 7 of them have been identified by the proposed pipeline. All the deletions, *69del*, *70del* and *144del*, are missed but adaptions where found as *I68M* and *H69G*. The substitutions *E484K*, *S494P* and *K1191N* were not identified. The *CDC* marked them as "*(*) detected in some sequences but not all*". Therefore, we treat it as 70% precision and 77% recall. Precision and recall values should only be seen as weak indicators for the predictive performance because the truth is not yet fully uncovered. Listed mutations might increase over time and false negatives and positives might change accordingly.

Lineage	Mutations
B.1.617.2	T19R, G142D, 156del, 157del, 158del, L452R, T478K, D614G, P681R, D950N
P.1	L18F, T20N, P26S, R190S, K417T, E484K, N501Y, D614G, H655Y, T1027I, V1176F
B.1.351	D80A, D215G, K417N, E484K, N501Y, D614G, A701V
B.1.1.7	I68M, H69G, Y144G, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H

TABLE I: Identified mutations by the pipeline.

V. Conclusion

The proposed pipeline hides the complexity of the underlying mathematical problem in a user-friendly way. Practitioners, such as geneticists, biologists, biochemists, virologists and epidemiologists, need easy access to complex analysis tools of high quality without the high barriers of understanding complex systems of equations. Each and every essential analysis step is supported by an intuitive and easy to understand graphical way. Most of the visualizations are interactive and provide the user with the opportunity to explore the data by his/her individual needs. Such an interactive framework will encourage users in their information search that hopefully increase their knowledge discovery. We proposed easy visual guidance in the data acquisition stage by providing overview of geolocations and meta-data distributions across regions, e.g. continents and countries. For the data analysis stage, especially the hierarchical clustering, we propose an interactive visualization which encourages the exploration of the clustering process. Using this toolbox we described how to interpret the clustering and how to draw conclusions from it. In the final step of the proposed analysis pipeline, we extract information from this clustering, visually analyze the sequences within the clusters and identify their specific mutation patterns. The proposed pipeline is a modular build and individual choices of models can be replaced. For a proof-of-concept we focus the application on the *spike glycoprotein* region within the SARS-CoV-2 genome for the variants of concern *B.1.1.7*, *B.1.351*, *B.1.617.2* and *P.1*. Other mutation groups can be analyzed as well by just a few clicks. The same holds true for other gene regions, such as the region for the *membrane glycoprotein*, *nucleocapsid phosphoprotein*, *envelope protein*, etc. For our initial evaluation we worked on a sub-sample of 3,506 genomes downloaded from the *GISAID Initiative*. It could be shown that we are able to detect mutations with reliable performance. In comparison to the mutations listed at the Centers for Disease Control and Prevention, we achieved precision recall values of 90% for *B.1.617.2*, 90.9% *P.1.*, 70% and 100% for *B.1.351* and 70% and 77% for *B.1.1.7*, respectively. Nonetheless, a detailed analysis showed obvious weaknesses in the detection of deletions. Overall this paper is not intended as a high-level bioinformatics paper. It is intended to build up a pipeline connecting existing tools to encourage practitioners in their research by providing interactive visualizations. The provided pipeline tries to bridge the gap of a complex problem task and intuitive knowledge discovery. We are aware that experts in the field of phylogenetics might (or even will) disagree with individual choices of models. But these individual choices can be replaced without conflicts in the described interactive framework. This work focuses on visualization, interaction and exploration. On that basis we believe that we could provide explainable solutions for users while preserving good predictive performance. The results of our proof-of-concept indicate that it could be sufficiently achieved in reference to the state-of-the-art. The proposed pipeline visualizes mutation variants and their dynamic, such as to characterize individual mutation positions. It could be assumed to be useful to identify new mutation variants. This provides the foundation of vaccine adaption for mutations because the mRNA vaccines technology works on this information within the described genomic regions.

References

[1] J. Hadfield, C. Megill, S.M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R.A. Neher Nextstrain: real-time tracking of pathogen evolution *Bioinformatics*, 34(23): 4121-4123, 2018

[2] I. Letunic, P. Bork Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation *Nucleic Acids Research*, 2021

[3] D.H. Huson, C. Scornavacca Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks *Systematic Biology*, 61(6): 10611067, 2012

[4] C.M. Zmasek, S.R. Eddy ATV: display and manipulation of annotated phylogenetic trees *Bioinformatics*, 17(4): 383384, 2001

[5] GISAID Initiative https://www.gisaid.org/

[6] A.C.E. Darling, B. Mau, F.R. Blattner, N.T. Perna Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements *Genome research*, 14(7): 1394-1403, 2004.

[7] T.H. Jukes, C.R. Cantor Evolution of Protein Molecules *Mammalian Protein Metabolism*, III: 21132, 1969.

[8] M. Kimura A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences *Journal of Molecular Evolution*, 16: 111120, 1980

[9] J. Felsenstein Evolutionary trees from DNA sequences: a maximum likelihood approach *Journal of Molecular Evolution*, 17: 368376, 1981

[10] M. Hasegawa, H. Kishino, T. Yano Dating of the human-ape splitting by a molecular clock of mitochondrial DNA *Journal of Molecular Evolution*, 22: 160-174, 1985

[11] M. Saitou, M. Nei The neighbor-joining method: a new method for reconstructing phylogenetic trees *Molecular Biology and Evolution*, 4: 406-425, 1987

[12] K.P. Schliep phangorn: phylogenetic analysis in R *Bioinformatics*, 27(4): 592-593, 2011

[13] L.T. Nguyen, H.A. Schmidt, A. von Haeseler, B.Q. Minh IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies *Molecular Biology and Evolution*, 32: 268-274, 2015

[14] A. Rambaut, E.C. Holmes, . OToole, V. Hill, J.T. McCrone, C. Ruis, L. du Plessis, O.G. Pybus A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology *Nature Microbiology*, 5: 1403-1407, 2020

[15] R. Durbin, S.R. Eddy, A. Krogh, G. Mitchison Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids *Cambridge University Press*, 1998

[16] J.A. Plante, Y. Liu, J. Liu, H. Xia, B.A. Johnson, K.G. Lokugamage, X. Zhang, A.E. Muruato, J. Zou, C.R. Fontes-Garfias, D. Mirchandani, D. Scharton, J.P. Bilello, Z.Ku, Z. An, B. Kalveram, A.N. Freiberg, V.D. Menachery, X. Xie, K.S. Plante, S.C. Weaver, P.-Y. Shi Spike mutation D614G alters SARS-CoV-2 fitness *Nature*, 592: 116121, 2021

[17] A. Larsson AliView: a fast and lightweight alignment viewer and editor for large datasets *Bioinformatics*, 30(22): 3276-3278, 2014

[18] Centers for Disease Control and Prevention (CDC), 20.05.2021 https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html

[19] Robert Koch-Institut (RKI), 20.05.2021 https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Virologische_Basisdaten.html