Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Phylogenicity of B.1.1.7 surface glycoprotein, novel distance function and first report of V90T missense mutation in SARS-CoV-2 surface glycoprotein

Done Stojanov

Faculty of Computer Science, Goce Delcev University of Stip, Republic of North MacedoniaKrste Misirkov, No.10 A F.O. Box 201, Stip 2000, Republic of North Macedonia

ARTICLE INFO

Keywords:
B.1.1.7
Novel distance function
V90T mutation
Escaping 2-51 neutralizing antibody

ABSTRACT

Phylogenicity of Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) B.1.1.7 surface glycoproteins reported from Europe to the National Center for Biotechnology Information (NCBI) virus database by the mid of April 2021 is analyzed. Novel function for computing phylogenetic distance is proposed for that purpose. Proposed distance function resulted in better-fitted clusters than Jaccard and Sorensen-Dice and accurate evolutionary links were predicted for B.1.1.7 spike variants. Most B.1.1.7 spike variants were linked to their likely direct predecessors at single amino acid change, that in many cases resulted in loss of the key mutations that are associated to the higher B.1.1.7 SARS-CoV-2 infectivity. There were also cases where second mutation was introduced to compensate for the missing mutation. Unreported V90T SARS-CoV-2 surface glycoprotein mutation was also identified, that contributes towards escaping 2-51 neutralizing antibody.

1. Introduction

In December 2020 Public Health England published a report named as “Variant of Concern 202012/01” that described preliminary findings for a newly detected variant of SARS-CoV-2 with alarming epidemiological properties (Public Health England, 2020). Emerging lineage, that was named B.1.1.7 (since 31 May 2021 labelled as Alpha by World Health Organization), emerged in the United Kingdom with two earliest samples collected on 20-Sept-2020 and 21-Sept-2020 in Kent and London respectively (Rambaut et al., 2020). B.1.1.7 variant has rapidly spread around the globe and up to date report is hosted on: https://cov-lineages.org/global_report/B.1.1.7.html. The rapid spread of B.1.1.7 variant can be explained by higher lineage infectivity (du Plessis et al., 2021) compared to preexisting lineage (43–90% higher, Davies et al., 2021a, 2021b), Davies et al. (2021a, 2021b) indicate that B.1.1.7 is not only more transmissible than preexisting lineage, but it can cause more severe illness. They estimated that the hazard of death associated with B.1.1.7 is 61% higher than pre-existing variants. B.1.1.7 variant defining mutations are found in: ORF1ab gene, S gene, ORF8 gene and N gene (Public Health England, 2020). Of all mutations, mutations found in the surface glycoprotein (S gene) are of the highest concern because surface glycoprotein (spike protein) mediates SARS-CoV-2 binding and entrance to the host cell (Shang et al., 2020). SARS-CoV-2 spike protein is synthesized as a 1273-amino acid trimer on the rough endoplasmic reticulum (Liu et al., 2020; Duan et al., 2020) and is composed of two subunits, S1 and S2 (Huang et al., 2020). Subunit S1 contains: N-terminal domain (NTD), receptor-binding domain (RBD) and two sub-domains (SD1 and SD2) of total size of 672 amino acids (residues 685–1273), Wrapp et al., 2020. Current research showed that spike protein mutations affect cell biding mechanism that results in higher infectivity (Starr et al., 2020). Some B.1.1.7 surface glycoprotein mutations, such as: N501Y, P681H, ΔHV [69–70], ΔY144 attracted special attention, as they have been associated to specific properties. N501Y mutation in the receptor-binding domain has been identified as the key mutation that increases virus binding affinity to human angiotensin-converting enzyme 2 (hACE2), Gu et al., 2020, P681H mutation, that is found next to the furin cleavage site, enhances virus transmission by facilitating spike protein conformational change (Lasek-Nesselquist et al., 2021). Hoffmann et al. (2020) found that cellular protease furin, cleavages surface glycoprotein at S1/S2 site, that is essential for cell fusion and entrance into the host cell. Multiple-Arginine S1/S2 cleavage site, which as such is not found in animal coronaviruses (Peacock et al., 2021), was reported as the key factor for fast transmission of B.1.1.7 SARS-CoV-2 variant. As for ΔHV [69–70] recurrent deletion, links to immune escape in immunocompromised patients were
made and it was proved that ΔHV [69–70] enhances viral infectivity in vitro (Volz et al., 2021). ΔY144 also contributes to immune escape, as Wang et al. (2021) found that ΔY144 plays pivotal role in resistance to NTD-directed monoclonal antibodies (mAbs).

1.1. Study aims

More than a hundred SARS-CoV-2 surface glycoprotein variants were computationally identified. NCBI Cobalt multiple alignment tool and local C# application developed in Visual Studio 2019 were used as basic means for computation. Exploring computed variants, rare missense mutations: A93Y and V90T were identified. A93Y is identified for the first time among the European population and V90T for the first time globally. Both mutations were found in highly-polymeric B.1.1.7 SARS-CoV-2 surface glycoprotein: QSQ87331.1. The impact of the novel V90T mutation on N-terminal domain biding affinity towards 2–51 neutralizing antibody was inspected by mCSM-PP2 web server. Estimated negative change in binding affinity of −0.040 kcal/mol indicated that V90T mutation contributes minimally towards escaping 2–51 neutralizing antibody.

B.1.1.7 spike variants were analyzed separately. It was found that most B.1.1.7 spike variants can be linked to their direct predecessors at single amino acid alteration. In many cases the alteration resulted in loss of key mutation that is associated to higher infectivity or many times B.1.1.7 SARS-CoV-2 evolved in less contagious forms. However, an exception to keep spike protein fit as much as possible was also identified, where another mutation was introduced to compensate for missing mutation. Isoleucine deletion at position 68 (ΔI68) is suggested to compensate for the missing Valine deletion at position 70 (ΔV70) in the commonly found ΔHV [69–70] deletion pair in one B.1.1.7 spike variant. Given that Valine and Isoleucine share similar physical-chemical properties and they are found next to each other in the N-terminal domain, ΔI68 may easily compensate for the missing ΔV70, without having to significantly alter residues' interactions and thereby preserve much of the native N-terminal domain conformation.

New formula for computing phylogenetic distance was proposed. Proposed formula can be used for reliable automated classification of any biological sequences. When tested on B.1.1.7 spike variants more reliable evolutionary paths were predicted for variants that experienced major changes compared to standard Jaccard and Sorensen-Dice metrics. Clusters’ quality was also improved most of the time.

2. Materials and methods

As far as of 12.04.2021, totally 897 records of SARS-CoV-2 surface glycoprotein variants were deposited to NCBI virus database (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#). All isolates were first preprocessed and then aligned to YP_009724390.1 referent SARS-CoV-2 Spike protein: Wuhan-Hu-1 complete genome (NCBI accession: NC_045512). Totally 28 records contained un-known amino acids and they were excluded from further analysis. The remaining 869 completely annotated and translated products of SARS-CoV-2 S genes were considered and aligned with NCBI Cobalt multiple alignment tool (https://www.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi; Papadopoulos and Agarwala, 2007), applying the following metric: opening gap penalty 2, extension gap penalty 1 and word size of 3.

Totally 124 SARS-CoV-2 surface glycoprotein variants were computed, Fig. 1. Group of spike proteins with the same combination of mutations profiles variant and at least one spike protein is required to report a new variant. Variants were computed by C# application, that I developed in Visual Studio 2019, that processed the list of mutations for each SARS-CoV-2 spike protein from Cobalt’s output and each time a new combination of mutations emerged, a new variant has been recorded, Fig. 1. If the list of mutations matched already computed variant, then the number of variant-matching spike proteins was only to be increased, Fig. 1. To record: variants’ mutations, the number of variant-matching spike proteins from NCBI virus database and additional data, such as geo-location, required advanced data structures to be engaged, such as: lists and dictionaries.

Since the start of pandemics, SARS-CoV-2 has evolved in several known VOCs (Variant of Concern) that served as a main drive for tracking virus polymorphisms and automated data classification. Software tool named PANGOLIN-Phylogenetic Assignment of Named Global Outbreak Lineages (https://cov-lineages.org/) designates isolates to the most likely lineage applying machine learning. On the other hand, Global Initiative on Sharing Avian Influenza Data-GISAID’s SARS-CoV-2 database (https://www.gisaid.org/hcov19-mutation-dashboard/) tracks complete and up to date list of worldwide SARS CoV-2 spike protein
mutations, organized by: most recent accession, isolate (also including geo-location data), frequency and date of collection.

Even designated under the same lineage, variants still profile specific polymorphisms. Deep analysis of mutual similarities and dissimilarities provided insights on transmission paths and dynamics of SARS-CoV-2 surface glycoprotein alteration. On the other hand, having computed variants profiling mutations, helped me track and report rare and unreported SARS-CoV-2 surface glycoprotein mutations.

Appendix A shows the complete list of computed SARS-CoV-2 surface glycoprotein variants. Missense mutations and deletions were found only among the European population. Each variant matched between 508 (max) and 1 (min) samples in the NCBI virus database. As expected, most European SARS-CoV-2 spike proteins include D614G mutation. This mutation was found as a single present mutation in 508 records (Europe top-ranked variant, Appendix A, Fig. 1). On the other hand, 103 variants included D614G in addition to another mutation and 20 variants did not include D614G at all. In terms of deletion, 40 variants included deletion and 84 variants contained only missense mutations. Most of the variants containing deletion (25) were designated to B.1.1.7 lineage.

2.1. Novel V90T SARS-CoV-2 spike protein mutation and rare C379F and A93Y mutations

From computed SARS-CoV-2 spike protein variants (Appendix A), I extracted 177 different mutations that circulate among the European population. Among them I found three extremely rare missense mutations: C379F, A93Y and V90T. GISAID's hCoV-19 spike mutations table (https://mendel.bi.a-star.edu.sg/METHODS/corona/current/MUTATION/HCoV-19_Human_2019_WuhanWIV04/hcov19_Spike_mutations_table.html) lists one worldwide occurrence of A93Y mutation, while V90T has not been reported as SARS-CoV-2 spike protein mutation so far. I found second worldwide occurrence of A93Y among the European population and this is the first report of V90T as SARS-CoV-2 surface glycoprotein mutation. None of the occurrences I report here have been evidenced in GISAID's hCoV-19 spike mutations table.

According to GISAID hCoV-19 Spike Glycoprotein mutation surveillance dashboard, missense mutation C379F occurred two times worldwide in spike protein from Switzerland: hCoV-19/Switzerland/ SO-ETHZ-490065/2020 and it was recently detected in spike protein from Indonesia: hCoV-19/Indonesia/JK-NIHRD-WGS07001/2021. The same missense mutation I found also in NCBI reported SARS-CoV-2 spike protein from France, accession: QJ772368.1, which was collected in March 2020. QJ772368.1 spike protein contains only: C379F and D614G mutations (Appendix A). On the other hand, missense mutations: A93Y and V90T were found in a highly-polymorphic B.1.1.7 surface glycoprotein variant, NCBI accession QSQ87331.1: H69-V70-P85-F86-N87-D88-G89-V90T-A93Y-D138HY144-D501YA5 70D614GP681H716S982AD1118H.

QSQ87331.1 B.1.1.7 spike protein was collected on 06.02.2021 in Spain. It contains all B.1.1.7 defining mutations plus major recurrent deletion ΔPFNDG [85 – 89] and three missense mutations: V90T, A93Y and D138H in the N-terminal domain. Of all B.1.1.7 spike variants I computed and there were 25 such, this is the most polymorphic B.1.1.7 surface glycoprotein that contains two missense mutations: V90T and A93Y that were identified for the first time among the European population. So far A93Y missense mutation is found only once worldwide in spike protein from Sri Lanka: hCoV-19/Sri Lanka/CMC11595/2020. Here I report the second worldwide occurrence of A93Y that I found in QSQ87331.1 spike protein. B.1.1.7 spike protein QSQ87331.1 also contains V90T missense mutation. Since there is no V90T spike mutation listed in GISAID's spike mutations table, this is the first international report of V90T as SARS-CoV-2 surface glycoprotein mutation.

Crystallographic structure of neutralizing antibody 2-51 in complex with SARS-CoV-2 spike N-terminal domain (NTD) (DOI: https://doi.org/10.2210/pdb7L2C/pdb, https://www.rcsb.org/structure/7l2c) was used to predict the effect of V90T missense mutation upon protein-to-protein interaction, Fig. 2. mCSM-PPI2 web server (Rodrigues et al., 2019; http://biosig.unimelb.edu.au/mcsmppi2/) was used to compute the change in binding affinity between SARS-CoV-2 spike N-terminal domain (NTD) and 2–51 neutralizing antibody upon Valine (non-polar, aliphatic) to Threonine (polar, non-charged) substitution at position 90 in N-terminal domain. The change in binding affinity is primarily computed upon residues' interactions which are affected upon missense mutation (different amino acids interact differently with surrounding residues, Fig. 2) that leads to protein conformation change inside the motif of interactions, that may increase or decrease protein binding affinity towards another protein, Fig. 2. Negative change of binding affinity $\Delta G = - 0.040 \text{ kcal/mol}$ for V90T missense mutation was computed what means that newly detected V90T missense mutation in the N-terminal domain of SARS-CoV-2 spike protein, even though minimally, contributes towards escaping 2–51 neutralizing antibody.

2.2. Phylogenicity of B.1.1.7 surface glycoprotein and novel distance function

Agglomerative clustering was performed on a subset of 24 unique B.1.1.7 spike protein variants, Fig. 3. The count of spike proteins per variant and location are listed in Table 1. The aim was to find clusters of closely related variants upon their mutations (Fig. 3) and thereby predict likely evolutionary links. Ward's method (Ward Jr, 1963) applied on distance matrix was used to compute dendrograms. A novel distance function for computing phylogenetic distance was proposed and evaluated relative to Jaccard (Jaccard, 1912; Tanimoto, 1958) and Sorenson-Dice (Dice, 1945; Sorensen, 1948; Nei and Li, 1979) distance functions. The distance function is defined in general and it can be applied to any two arbitrary sets. When applied to the subset of B.1.1.7 spike variants on Fig. 3, proposed distance function resulted in better-fitted clusters (for higher number of clusters K) than Jaccard and Sorenson-Dice and highly–polymorphic variants were accurately linked to corresponding clusters. Multidimensional scaling was used to visualize clusters' structure in 2D space. The complete analysis was performed in Python Software Foundation for Python 3.9 (Python 3.9).

Even though 25 B.1.1.7 spike protein variants were computed, only 24 of them were considered, Fig. 3. Since long recurrent deletions in the N-terminal domain (NTD) of SARS-CoV-2 spike protein are specific to long-term infections in immunocompromised patients (Avanzato et al., 2020; Aydillo et al., 2020; Choi et al., 2020; Hensley et al., 2021), previously analyzed B.1.1.7 spike protein QSQ87331.1 is supposed to be found in such patient. In terms of computation, QSQ87331.1 is considered as an obvious outlier. On the other hand, there is no abrupt change in the rate of mutations for the remaining 24 B.1.1.7 spike variants on Fig. 3 and an assumption that those variants were isolated from non-immunocompromised patients was made. Clues towards the dynamics and type of B.1.1.7 spike protein alteration are aimed to be identified, as SARS-CoV-2 was transmitted among non-immunocompromised patients in Europe, based on the available data (Fig. 3).

Comparing B.1.1.7 spike protein variants on Fig. 3 to B.1.1.7 spike protein defining mutations – defining deletions: ΔH69, ΔL70, ΔY144/ΔY145 plus defining amino acid substitutions: N501Y, A570D, D614G, P681H, T716I, S982A, D1118H (Galloway et al., 2021; Mulk et al., 2021), two types of events can be observed in 21 out of 24 variants:

1. Deletion of defining amino acid substitution
2. Non-defining amino acid substitution

Histidine deletion (ΔH69) and Asparagine to Tyrosine substitution (N501Y) are found to be the most stable B.1.1.7 surface glycoprotein mutations (Fig. 3). Variants V1(V2) experience no additional mutation, Fig. 3. As Fig. 3 shows, non-synonymous single nucleotide polymorphisms (SNPs) at codons for defining amino acid substitutions in B.1.1.7 S genes did not result in random protein mutations, but on the
contrary, for more than half variants on Fig. 3, the amino acid in the referent Wuhan SARS-CoV-2 spike protein (YP_009724390.1) was reversed—an event that is regarded as deletion of defining amino acid substitution. In 13 out of 24 variants: V5, V [7–12], V16, V [20–24] (Fig. 3), deletion of defining amino acid substitution can be observed. On the other hand, non-defining amino acid substitutions also occurred, Fig. 3. Non-defining amino acid substitution and no deletion of defining substitution can be observed in 8 variants: V4, V6, V [13–15], V [17–19] (Fig. 3). Some variants: V11, V16, V [20–24] included both events (1 and 2) (Fig. 3). Uncommon alterations: ∆I68 and V70I were found in variant V3 (Fig. 3) and they are discussed later in the paper.

Here I consider changes at protein level, given that B.1.1.7 spike proteins underwent most of the time edit events of type: (1) and (2), as SARS-CoV-2 was transmitted among non-immunocompromised patients. Following these suggestions, cluster analysis towards identifying edit history is performed. While “absolute edit history” connects variants to their common ancestor (defined by the list of B.1.1.7 spike protein defining mutations), “relative edit history” and especially minimal relative edit history between variants, boosted with date of collection and geo-location data may reveal likely changes in strains that occurred as the virus was transmitted form one host to another.

The first step of computational analysis is to define function for computing phylogenetic distance between variants. Given that variants are expressed as sets of mutations at protein level (Fig. 3), applicable distance functions compute phylogenetic distance upon dissimilarity between sets. The way dissimilarity is computed may impact classification output or different clusters may be obtained given that different distance functions are applied. However, regardless of the distance function being used, a real number in the range between 0 and 1 is always returned, such 1 is returned for completely different sets and 0 for identical sets. Even though Jaccard (Jaccard, 1912; Tanimoto, 1958) distance function and Sorensen-Dice (Dice, 1945; Sorensen, 1948; Nei and Li, 1979) distance function are the defaults when it comes to measure phylogenetic distance, here I propose a novel distance function that can be used to measure the dissimilarity between any two random sets in general and thereby it can be also applied to compute phylogenetic distance between variants on Fig. 3. Proposed distance function can be formulated as:

![Fig. 2. Protein-to-Protein interactions (wild type Valine vs mutant Thrreonine at 90 in NTD).](image_url)

![Fig. 3. B.1.1.7 spike protein variants.](image_url)
Table 1

| Variant | Count | Geo-location |
|---------|-------|--------------|
| V1      | 35    | Spain (35)   |
| V2      | 21    | Austria (20), Germany (1) |
| V3      | 7     | Italy (7)    |
| V4      | 5     | Spain (5)    |
| V5      | 5     | Spain (5)    |
| V6      | 2     | Austria (2)  |
| V7      | 2     | Spain (2)    |
| V8      | 2     | Spain (2)    |
| V9      | 2     | Spain (2)    |
| V10     | 2     | Spain (2)    |
| V11     | 2     | Spain (2)    |
| V12     | 2     | Spain (2)    |
| V13     | 2     | Spain (2)    |
| V14     | 1     | Finland (1)  |
| V15     | 1     | Spain (1)    |
| V16     | 1     | Spain (1)    |
| V17     | 1     | Spain (1)    |
| V18     | 1     | Spain (1)    |
| V19     | 1     | Spain (1)    |
| V20     | 1     | Spain (1)    |
| V21     | 1     | Spain (1)    |
| V22     | 1     | Spain (1)    |
| V23     | 1     | Spain (1)    |
| V24     | 1     | Spain (1)    |

\[
d(A, B) = \frac{|A \backslash B| |A| + |B \backslash A| |B|}{|A|^2 + |B|^2} (1)
\]

such as \(d(A, B)\) denotes distance between sets \(A\) and \(B\), \(|\cdot|\) denotes cardinality of set, \(A'\) is the relative complement of \(B\) in \(A\) and \(B'\) is the relative complement of \(A\) in \(B\). Under set cardinality \(|\cdot|\) we assume the number of elements contained in the set, \(A'\) denotes elements contained in set \(B\) but not in set \(A\). The sum of cardinalities of sets \(A'\) and \(B'\) multiplied by cardinality of sets \(A\) and \(B\) and then divided by the sum of squared cardinalities of sets \(A\) and \(B\), is how the distance is computed following proposed metric – Eq. (1).

In Appendix B, I prove that proposed distance function according Eq. (1) satisfies properties: P1, P2, P3 and P4. Note that properties where equivalence: \(\Leftrightarrow\) or “if and only if” (abbreviated by “iff”) is suggested, are proved in “both directions”.

P1: \(0 \leq d(A, B) \leq 1\)
P2: \(d(A, B) = 0 \Leftrightarrow A = B\) (equivalent sets).
P3: \(d(A, B) = 1 \Leftrightarrow (A \cap B) = \emptyset\) (disjoint sets).
P4: \(d(A, B) = d(B, A)\) (symmetry).

Even though distance functions share some common properties, different distance may be computed for the same samples by different distance function. That will be demonstrated for variants V1 and V3 on Fig. 3, given that: Jaccard Eq. (2), Sorensen-Dice Eq. (3) and proposed Eq. (1) are applied. Jaccard distance function is defined by Eq. (2) and the distance between sets is computed upon cardinality of intersection and union. Cardinality of intersection is also considered in Sorensen-Dice distance function Eq. (3), but there instead of cardinality of union, the sum of cardinalities of sets is used.

\[
d(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} (2)
\]

\[
d(A, B) = 1 - \frac{2|A \cap B|}{|A| + |B|} (3)
\]

Given that variants V1 and V3 on Fig. 3 are represented by their respective mutations on protein level, one gets:

\[V1 = \{H69-, V70-, Y144-, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H\}.\]

To compute Jaccard distance between variants V1 and V3, intersection and union of mutations for V1 and V3 needs to be computed, such as common mutations are assigned to the intersection and common plus different mutations to the union:

\[V1 \cap V3 = \{H69-, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H\}.\]

\[V1 \cup V3 = \{H69-, H69-, V70-, V70I, Y144-, Y144-, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H\}.\]

Since \(|V1 \cap V3| = 8\) and \(|V1 \cup V3| = 13\) Jaccard \(d(V1, V3)\) distance equals:

\[d(V1, V3) = 1 - \frac{|V1 \cap V3|}{|V1 \cup V3|} = 1 - \frac{8}{13} = 0.846\]

On the other hand, having: \(|V1| = 10, |V3| = 11\) and \(|V1 \cap V3| = 8\), Sorensen-Dice \(d(V1, V3)\) distance equals:

\[d(V1, V3) = 1 - \frac{2|V1 \cap V3|}{|V1| + |V3|} = 1 - \frac{2 \times 8}{10 + 11} = 0.2381\]

To apply proposed Eq. (1), set-theoretic difference of variant V3 in V1 and variant V1 in V3 needs to be computed, such as \(V1 \setminus V3\) equals all mutations found in V1 but not in V3 and \(V3 \setminus V1\) equals all mutations found in V3 but not in V1:

\[V1 \setminus V3 = \{V70-, Y144-\}\]

\[V3 \setminus V1 = \{H69-, V70I, Y145-\}.
\]

\[V1 \Cap V3 = \{H68-, -V70, V145-\}.
\]

Given that: \(|V1 \setminus V3| = 2, |V3 \setminus V1| = 3\) and \(|V1| = 10, |V3| = 11\), one can easily compute phylogenetic distance between variants V1 and V3 according Eq. (1):

\[d(V1, V3) = \sqrt{\frac{|V1 \setminus V3| + |V3 \setminus V1|}{|V1|^2 + |V3|^2}} = \sqrt{\frac{2 \times 10 + 3 \times 11}{100 + 121}} = 0.2398\]

Since different distances may impact how variants are linked together, different structures of clusters may be computed given that the same clustering algorithm is applied to distance matrix \(D\) computed by different distance metric. Distance matrix \(D\) here contains all pairwise distances \(d(V_i, V_j)\) between variants on Fig. 3 and \(D = [d(V_i, V_j)]_{24 \times 24}\) for Eq. (1), Jaccard Eq. (2) and Sorensen-Dice Eq. (3) were computed. Appendix C lists condensed distance matrix \(D\) for variants on Fig. 3 applying Eq. (1). Condensed distance matrix is a vector form representation of a symmetrical distance matrix \(D\) that contains \(D\)’s upper triangle.

Ward’s method (Ward Jr, 1963) was used as clustering algorithm. Ward’s algorithm iteratively combines clusters together, based on criterion for minimum increase of the error sum-of-squares. Given that the distance between clusters \(C_i\) and \(C_j\) is computed as: \(d(C_i, C_j) = \text{ESS}(C_i \cup C_j) - \text{ESS}(C_i) - \text{ESS}(C_j)\), such as: \(\text{ESS}(C_j)\) denotes the error sum-of-squares for cluster \(C_j\): \(\text{ESS}(C_j) = \sum_{x \in C_j} (x - m_{C_j})^2\) and \(m_{C_j}\) is cluster \(C_j\) centroid or the simple mean of all samples \(x\) contained in cluster \(C_j\):

\[m_{C_j} = \frac{\sum_{x \in C_j} x}{|C_j|}\]

The centroids for each cluster \(C_j\) are computed by the squared difference of each sample \(x\) from the centroid \(m_{C_j}\). Initial set of samples \(x\) is considered as a separate cluster and clusters are merged together based on \(d \rightarrow \min\) criterion, until one cluster is obtained in the final phase of the process. In other words, given that there are \(k\) candidate clusters: \(C = \{C_1, C_2, ..., C_k\}\) that are to be merged with cluster \(C_l\) at some phase of the clustering, Ward’s algorithm will select cluster \(C_l\) from the list of candidates in \(C\), such as: \(d(C_l, C_j) = \min\{d(C_l, C_1), d(C_l, C_2), ..., d(C_l, C_k)\}\) or minimum increase of the error sum-of-squares is attained. Joining clusters together...
upon intra-cluster disturbance minimization criteria, usually results in best-differentiated clusters, what was the main reason why Ward’s method was selected over the other available linkage methods in Python. Given that clusters $C_i$ and $C_j$ were already joined together into one resulting cluster: $C_{i,j} = C_i \cup C_j$, cluster $C_k$ from the current state of disjoint clusters for which Eq. (4) is minimized, is merged to $C_{i,j}$: 

$$d(C_{i,j}, C_k) = \sqrt{\frac{|C_i||C_k|}{|C_i| + |C_k|} d(C_i, C_k)^2 + \frac{|C_j||C_k|}{|C_j| + |C_k|} d(C_j, C_k)^2 - \frac{|C_i|}{|C_i| + |C_j|} d(C_i, C_j)^2 \text{Eq.}(4).$$

3. Results

Ward’s algorithm was applied to distance matrix $D$ that was computed for B.1.1.7 spike protein variants on Fig. 3 by: Eq. (1), Jaccard Eq. (2) and Sorensen-Dice Eq. (3) distance functions and dendrograms on Fig. 4 were obtained.

To estimate the impact of different distance functions over the quality of clustering, silhouette scores for $K = 2, 3, 4, 8, 10$ clusters were computed, based on dendrograms on Fig. 4. Silhouette coefficient $s(V_i)$ measures how well variant $V_i$ fits its own cluster $C$ compared to the nearest cluster $C_0$, such that: $C$ and $C_0$ are two different clusters. Two parameters $a$: intra-cluster mean distance and $b$: nearest-cluster mean distance relative to $V_i$ define $s(V_i)$: $s(V_i) = \frac{b - a}{d_{max}}$ such that: $a = \frac{1}{|C|} \sum_{V_j \in C} d(V_i, V_j)$ and $b = \frac{1}{|C_i|} \sum_{V_j \in C_i} d(V_i, V_j)$. Given that parameter $a$ measures $V_i$’s mean distortion inside its own cluster $C$ and $b$ measures $V_i$’s mean distance to the nearest cluster $C_0$, $V_i$ is considered to be well-fitted to its own cluster $C$ for low intra-cluster mean distance ($a$) and high mean distance to the nearest cluster ($b$). A real number in the range: $-1$ to $+1$ is returned for the silhouette coefficient, such as: $s(V_i) \approx +1$ denotes that $V_i$ fits well its own cluster $C$, $s(V_i) \approx 0$ that $V_i$ is on the border between its own and neighboring cluster and $s(V_i) \approx -1$ that $V_i$ is assigned to the wrong cluster or $V_i$ fits better neighboring cluster $C_0$ than its own cluster $C$. Silhouette score for particular classification equals the mean silhouette coefficient for all samples, such as: positive silhouette score indicates differentiated clusters, zero that clusters touch each other and negative that clusters overlap. Silhouette score was taken as a measure for the quality of clusters obtained by Ward’s algorithm on three different distance functions: Eq. (1), Jaccard Eq. (2) and Sorensen-Dice Eq. (3) for B.1.1.7 spike protein variants on Fig. 3.

Three different dendrograms were computed for three different distance functions: Eq. (1), Jaccard Eq. (2) and Sorensen-Dice Eq. (3), Fig. (4, 5). Clusters’ structure for $K = 2, 3, 4, 8, 10$ clusters is found by intersections of horizontal lines on various levels on Fig. 5. Table 2 lists clusters’ structure for each $K$, for each dendrogram and given that Jaccard and Sorensen-Dice produced same structures of clusters for all $K$’s, Eq. (1) did not, Fig. (4, 5).

Multidimensional scaling was used to visualize clusters in $\mathbb{R}^2$ space, Fig. 6. In general, multidimensional scaling uses pairwise distances from distance matrix to compute coordinates for each sample in lower $\mathbb{R}^N$ space ($N$ usually equals 2 or 3), such that distance between samples in $\mathbb{R}^N$ space approximates pairwise distance from distance matrix. Given that transformation error is controlled by $\text{eps}$ parameter in Python, Fig. 6 shows clusters’ structure for Ward’s output upon multidimensional scaling, for: $N = 2$, $K = 8$ clusters and $\text{eps} = 1e-12$. Variants assigned to the same cluster are marked with the same color on Fig. 6. For higher number of clusters $K$, $K = 8$ and $K = 10$, proposed distance function Eq. (1) resulted in different clusters than Jaccard (Sorensen-Dice), Table 2, Fig. 5. Clusters that were differently computed from Jaccard (Sorensen-Dice) are marked in red color in Table 2. Silhouette score analysis – Table 3, Fig. 7 shows that Eq. (1) improved the quality of clusters for higher number of clusters $K$ (Table 3, Fig. 7) and more accurate links were predicted for variants that are on the edge of classification, compared to Jaccard and Sorensen-Dice.

For $K = 8$ clusters and Jaccard (Sorensen-Dice) distance matrix, Ward’s algorithm clustered variants: V18 and V22 together, while for proposed Eq. (1) they were classified separately, Table 2, Fig. (5, 6). Variant V18 was clustered with variants: V12, V19, V15, V1, V13, while variant V22 formed another cluster with: V9, V23, Table 2, Fig. (5, 6). Linkage change for V18(V22) due to Eq. (1) resulted in better-fitted clusters compared to Jaccard and Sorensen-Dice, Table 3, Fig. 7. Silhouette score of 0.3348169603112492 was computed for $K = 8$ clusters given that distance matrix was computed by Eq. (1), while silhouette scores given that Jaccard and Sorensen-Dice distance
metaGene 30 (2021) 100967

7

D. Stojanov

Fig. 5. Clusters’ structure.

clusters was improved. Two edit events: L18F non-defining substitution and deletion of defining S982A substitution are required to transform variant V1 into V18 (Fig. 3). Since Eq. (1) implied variants at the best edit distance to be classified together (V22 and V9 in cluster C6; V18 and V1 in cluster C8, Table 2, Fig. (5, 6)), while by Jaccard and Sorensen-Dice they were not, Eq. (1) is suggested as an improved option for predicting more accurate structural and evolutionary links between samples for higher number of clusters K, where usually other distance metrics fail.

Based on computed silhouette scores in Table 3 for Ward’s clustering output for: Eq. (1), Jaccard Eq. (2), Sorensen-Dice Eq. (3) distance matrix, for K = 2, 3, 4, 8, 10 clusters, one can easily find that Eq. (1) generated better-fitted clusters than Jaccard in all cases, while compared to Sorensen-Dice, better-fitted clusters were computed for higher values of K: K = 8 and K = 10, Table 3, Fig. 7. For K = 10 clusters, silhouette score for Eq. (1) equals 0.3249632576792248, while for Jaccard and Sorensen-Dice: 0.29527962158658155 and 0.3146922924108971 respectively, Table 3, Fig. 7. The same discussion can be applied for K = 8 clusters, where silhouette score of 0.30333844570840973 and 0.3249632576792248 for Jaccard and Sorensen-Dice, Table 3, Fig. 7. For lower number of clusters K = 2, 3, 4, Eq. (1) resulted in a bit worse-fitted clusters than Sorensen-Dice, Table 3, Fig. 7.
Fig. 6. Multidimensional scaling for $K = 8$ clusters.

Table 3
Silhouette scores.

| Number of clusters K | Jaccard     | Sørensen-Dice | Eq. (1)     |
|----------------------|-------------|---------------|-------------|
| K = 2                | 0.3514892909348634 | 0.39391740091699345 | 0.38358135735917775 |
| K = 3                | 0.30097976038432583 | 0.3395591191549989 | 0.3368693854380697 |
| K = 4                | 0.277827453162159  | 0.30115925901503604 | 0.289477743151141 |
| K = 8                | 0.3033844570840973 | 0.3249632576792248 | 0.3348169603112492 |
| K = 10               | 0.29527962158658155 | 0.3146922924108971 | 0.32145288135845923 |

Fig. 7. Silhouette scores: Jaccard, Sørensen-Dice and Eq. (1).
Fig. 7. Since Eq. (1) outperformed Jaccard in all cases and Sorensen-Dice for higher number of clusters K (Table 3, Fig. 7), Eq. (1) is suggested as better option than Jaccard and Sorensen-Dice for accurate prediction of closely-tied phylogenetic relations.

4. Discussion

Given that B.1.1.7 spike protein underwent most of the time: deletion of defining amino acid substitution and non-defining amino acid substitution – Fig. 3, in this section I try to find likely phylogenetic links for B.1.1.7 spike protein variants on Fig. 3, based on Ward’s classification output for Eq. (1) and K = 8 clusters (Table 2, Fig. (5, 6)). Variants’ structure (Fig. 3) and date/place of collection are also considered when suggesting likely linkage.

Highly-polymorphic variants V3 and V20 were accurately predicted as single-variant clusters, Table 2, Fig. (5, 6). In particular, variant V3 replaced ΔV70 with V70I and most likely the absence of ΔV70 in the commonly found B.1.1.7 ΔHV {69,70} spike deletion pair was compensated by deletion of Isoleucine at the nearby position 68: ΔI68, Fig. 3. I hypothesize this, because both Valine and Isoleucine are non-polar aliphatic amino acids with similar physical-chemical properties (most hydrophobic amino acids, no hydrogen donor or acceptor atom in side chains, very close: pKₐ, pKₐ', pi, ... etc.) and next to each other, so the one could easily compensate for the missing deleterious effect of the other. In the case of variant V3, ΔI68 probably compensates for the missing ΔV70, Fig. 3.

On the other hand, variant V20 profiles most deletions of defining amino acid substitutions: A570D, D614G and T716I, Fig. 3. As mentioned before, deletion of defining substitution means that amino acids: {D(570), G(614), I(716)} were reversed to amino acids: {A(570), D(614), T(716)}, the same that are contained in Wuhan SARS-CoV-2 referent spike protein (YP_009724390.1). As variant V20 lost D614G mutation which is the key mutation that is associated to enhanced SARS-CoV-2 replication in human lung epithelial cells and primary human airway tissues, reduced infectivity may be associated to V20. Reduced infectivity of variant V20 is also affirmed by the fact that only one spike protein (QTC10682.1) matched variant V20, Table 1.

Cluster C2 consists of variants: V14, V2, V6, Fig. (5, 6), Table 2. Earliest reported variant V14 B.1.1.7 spike protein (QTH25075.1) was collected on 22.12.2020 in Germany, while all remaining 20 V2 spike proteins were collected in Austria and for the most of them 02.2021 was reported as collection date. V6 spike proteins: QST76474.1 and QST76510.1 were collected on 17.02.2021 and 19.02.2021 in Austria, while V14 spike protein QSS48541 was collected on 18.12.2020 in Finland. Variants assigned to cluster C2 experience no deletion of defining amino acid substitution and they are profiled by ΔY145 instead of ΔY144 deletion (Fig. 3).

In this case, variant V2 is suggested as likely parental strain for variant V6, as variant V6 might have been transformed from variant V2 upon S155G substitution (Fig. 3). Given that the earliest V2 spike protein (QTH25075.1) was collected in Germany and all others were later collected in Austria, variant V2 might have been transmitted from Germany in Austria. Due to ΔY145 deletion in V14 (Fig. 3), V14 was also linked to V2 and V6, Fig. (5, 6).

Variants: V17, V4 and V11 were linked together to form cluster C4, Fig. (5, 6), Table 2. Spike protein QTC10874.1 is the earliest collected V4 sample on 14.01.2021 in Spain. Earliest collected V11 and V17 B.1.1.7 spike protein are: QTC10766.1 and QTC10658.1, that were collected on 19.01.2021 and 21.01.2021, also in Spain. For this cluster, variant V4 is suggested as likely parental strain for variant V11 (Fig. 3). Variants V11 is predicted to emerged from V4 upon deletion of defining P681H substitution (Fig. 3). The absence of P681H mutation in this case can be associated to decreased SARS-CoV-2 binding affinity towards host cell, what is expected to result in reduced infectivity and mild up to moderate clinical presentation.

Four variants: V24, V21, V8 and V16 form cluster C5, Fig. (5, 6), Table 2. There are two V8 B.1.1.7 spike proteins: QTC10250.1 and QTC10982.1 collected on 13.01.2021 and 14.01.2021 in Spain and one B.1.1.7 spike protein per variants V16, V21 and V24: QTC10634.1, QTC10790.1 and QTC10802.1 collected on 25.01, 19.01 and 19.01.2021 also in Spain. Variants assigned to cluster C5 are missing A570D defining mutation, Fig. 3. Since A570D mutation makes close to open spike conformational update (Warwicker, 2021), no amplification of infectivity than already introduced by D614G mutation (Zhang et al., 2020) can be attached to variants: V24, V21, V8 and V16 (Fig. 3).

In this case variant V8 is suggested as likely parental strain for variants: V16, V21 and V24, such as: V16 is predicted to emerged from V8 upon S943I substitution in V8, V21 upon E96D substitution in V8 and V24 upon A653V substitution in V8 (Fig. 3).

Ward’s algorithm applied on Eq. (1) classified variants: V22, V9, V23 together, Fig. (5, 6), Table 2. Variants V22, V9 and V23 are missing P681H defining substitution, Fig. 3. Given that P681H mutation enhances SARS-CoV-2 binding affinity towards the cell that results in increased infectivity, the absence of P681H mutation for V22, V9 and V23 may result in reduced infectivity. Following variants structure and collection dates, variant V9 is about to be suggested as likely predecessor for variants: V22 and V23 (Fig. 3), V9 spike proteins: QTC11006.1 and QTC10454.1 were collected on 13.01.2021 in Spain, while V22 and V23 spike proteins: QTC10598 and QTC10538.1 were collected on 18.01 and 13.01.2021 at the same geo-location. Suggestion that variant V23 might have been transformed from V9 upon E96D substitution is made (Fig. 3), while V22 can be also linked to V9, but two alterations must have happened: L18F non-defining substitution and deletion of S982A defining substitution (Fig. 3).

Although V5 was found nearest to V10 than V7, when it comes to cluster C7 (Fig. 5, 6), Table 2), it is equally likely that V10 emerged either from V5 or V7 (Fig. 3). Given that V10 emerged from V5, then D1118H deletion must had occurred in V5 (Fig. 3). On the other hand, given that V10 emerged from V7, then S982A deletion must had occurred in V7 (Fig. 3). As reversal of defining amino acid substitution upon its deletion is considered unlikely to happen, variants V5 and V7 could not have emerged one from other, Fig. 3. Earliest reported V5 spike proteins: QTC10394.1 and QTC10466.1, earliest reported V7 spike protein QTC10622.1 and V10 spike protein QTC10946.1 were collected on 13.01.2021 in Spain.

The largest cluster C8 contains six variants: V18, V12, V19, V15, V1, V13, Fig. (5, 6), Table 2. For this cluster, V1 is suggested as likely parental strain for variants: V12, V13, V15 and V19. Earliest reported V1 spike proteins: QTC10190, QTC10202 and QTC10226 were collected on 09.01.2021, V12 spike proteins: QTC10910.1 and QTC10490.1 were collected on 13.01 and 14.01.2021, V13 spike proteins: QTC10934.1 and QTC10574.1 were collected on 13.01.2021, V15 spike protein QTC10262 on 19.01.2021 and V19 spike protein QTC10442 on 14.01.2021 in Spain. Here I suggest that variant V12 emerged upon D614G deletion in V1, variant V13 upon E96D substitution in V1, variant V15 upon P909S substitution in V1 and variant V19 upon L5F substitution in N-terminal domain of V1, Fig. 3. Due to Eq. (1) a bit more distant linkage for variant V18 was also correctly predicted by assigning V18 to cluster C8, such as variant V18 might have evolved (directly or indirectly) from V1 upon substitutions: L18F and P812L, Fig. 3. Variant V18 spike protein QTC10862 was collected on 18.01.2021 in Spain. Reduced infectivity may be associated to V12, as variant V12 lost D614G mutation, Fig. 3.

Given that reliable evolutionary paths were traced for the majority of B.1.1.7 spike variants that underwent single amino acid alteration, suggestions for direct predecessors could not have been identified for variants that underwent severe alteration. One of the possibilities is that their parental strains were not reported to the NCBI virus database and therefore they could not have been linked properly. In this category one can enumerate variants: V3, V18, V20 and V22. Severe polymorphism of variant V20 and specific alterations found in variant V3, suggested these variants as outliers since the very beginning of the study and no significant cluster association for these variants could have been identified.
On the other hand, variants V18 and V22 were linked to more specific clusters, but likely direct predecessors for V18(V22) could not have been also identified.

Other weakness of the study is that proposed phylogenetic formula Eq. (1) underperformed Sorensen-Dice in clusters' quality for low number of clusters K. Ward's algorithm applied on Sorensen-Dice metric for B.1.1.7 spike variants generated more compact classification for K = 2, 3, 4 clusters than Eq. (1) (Table 3, Fig. 7), that suggests Eq. (1) as improved option only for high number of clusters K.

5. Conclusion

As B.1.1.7 SARS-CoV-2 evolved in many other forms, B.1.1.7 spike protein lost some of the key amino acid mutations that are associated to the higher B.1.1.7 infectivity or less contagious B.1.1.7 variants compared to their likely predecessors were identified. Most variants were linked to their likely direct predecessors at single amino acid change or only one codon in B.1.1.7 S gene was found to be affected. Even though this change in some variants resulted in loss of mutation that is associated to the higher B.1.1.7 infectivity, in some cases a novel mutation was introduced to compensate for the missing mutation. Such is the case of variant V3, where Isoleucine deletion at position 68 – ∆I68 is supposed to be compensation for the missing deleterious effect of Valine at position 70 – ∆V70 in the N-terminal domain.

Unreported V90T missense mutation in the N-terminal domain of SARS-CoV-2 spike protein was identified. V90T mutation showed positive affirmation towards escaping 2–51 neutralizing antibody. Novel function for computing phylogenetic distance was also proposed, that did better structural and evolutionary links predictions than Jaccard and Sorensen-Dice. Silhouette score analysis demonstrated that proposed distance function resulted in better-fitted clusters for higher number of clusters, what makes it suitable for more accurate predictions of closely-tied relations.

As for further research, proposed phylogenetic formula will be applied and evaluated to other biological sequences. In addition to popular clustering algorithms, more reliable evolutionary trajectories are expected to be automatically predicted than custom metrics, that is expected to bring better understanding of some evolutionary changes to the scientific community. Predicting automatically the accurate history of alterations is crucial for building reliable epidemiological models, that may provide better understanding of the phylogenicity of many pathogens, including highly alarming SARS-CoV-2, that is of the highest priority nowadays.

Declaration of Competing Interest

None.

Appendix A. Computed SARS-CoV-2 surface glycoprotein variants

D614G, Hits: 508.
H69-V70-Y144-N501YA570DD614GP681HT716IS982AD1118H, Hits: 35.
H69-V70-Y145-N501YA570DD614GP681HT716IS982AD1118H, Hits: 21.
D614GD839Y, Hits: 16.
A222VD614G, Hits: 11.
D614GA1020V, Hits: 7.
I68-H69-V70Y145-N501YA570DD614GP681HT716IS982AD1118H, Hits: 7.
D614GM740I, Hits: 7.
LSFD614G, Hits: 6.
D80YN439KD614G, Hits: 5.
A845S, Hits: 5.
H69-V70-Y144-L216FN501YA570DD614GA653VP681HT716IS982AD1118H, Hits: 5.
H69-V70-Y144-N501YA570DD614GP681HT716ID1118H, Hits: 5.
H49YQ675-T676-Q677-T678-N679-, Hits: 4.
H49YD614G, Hits: 4.
D614GT723I, Hits: 3.
TS47ID614G, Hits: 3.
D614GA1078V, Hits: 3.
D614GV687I, Hits: 3.
H69-V70-Y145-S155GN501YA570DD614GP681HT716IS982AD1118H, Hits: 2.
S98FD614G, Hits: 2.
H49Y, Hits: 2.
D614GT676IP1162L, Hits: 2.
S477ND614G, Hits: 2.
L18FD614GA1020V, Hits: 2.
I197V, Hits: 2.
M153ID614G, Hits: 2.
L18FT20NP26SD138YR190SK417TE484KN501YD614GA701V, Hits: 2.
T76ID614G, Hits: 2.
A243-L244-D614G, Hits: 2.
D614GA626S, Hits: 2.
H245YD614G, Hits: 2.
A372VD614G, Hits: 2.
H69-V70-Y144-N501YA570DD614GP681HT716IS982AD1118H, Hits: 2.
H69-V70-Y144-N501YD614GP681HT716IS982AD1118H, Hits: 2.
S12D614GD627A, Hits: 2.
L18FP139-F140-L141-G142-L244-A222VD614G, Hits: 2.
H69-V70-Y144-N501YA570DD614GT716IS982AD1118H, Hits: 2.
H69-V70-Y144-N501YA570DD614GP681HT716I, Hits: 2.
H69-V70-Y144-N501YA570DD614GA653VT716IS982AD1118H, Hits: 2.
T20IV503ID614GT1117I, Hits: 2.
L18FA222VD614G, Hits: 2.
H69-V70-Y144-N501YA570DP681HT716IS982AD1118H, Hits: 2.
H69-V70-E96DY144-N501YA570DD614GP681HT716IS982AD1118H, Hits: 2.
H69-V70-N439KD614G, Hits: 1.
F565LD614QV772SF888L, Hits: 1.
P26LH69-V70-L189FN439KD614G, Hits: 1.
LBWH69-V70-N439KD614GQ926H, Hits: 1.
S477ND614GD950AIS1196F, Hits: 1.
P9LS477NE484KD614G, Hits: 1.
T259ID614GDI1118Y, Hits: 1.
I68-H69-V70-S71-G72-T73-N74-G75-T76-T676-Q677-T678-N679-S680-S813I, Hits: 1.
H69-V70-Y145-N501YA570DD614GP681HT716IA831VS982AD1118H, Hits: 1.
F59YP384LD614GP812S, Hits: 1.
T21, Hits: 1.
D80AD215GL242-A243-L244-K417NE484KN501YD614GA701VQ836L, Hits: 1.
K558R, Hits: 1.
S254FD614G, Hits: 1.
N188DD614G, Hits: 1.
D614GG769V, Hits: 1.
V1122L, Hits: 1.
D614GH101YK1191N, Hits: 1.
T393PD614G, Hits: 1.
D614GA892S, Hits: 1.
D614GY789D, Hits: 1.
D614GB780V, Hits: 1.
T299ID614G, Hits: 1.
I68-H69-V70IN439KD614GD1146Y, Hits: 1.
T1009I, Hits: 1.
Q5RA67V168-H69-V70IY145-M153TV401LE484KD614GQ677HF888L, Hits: 1.
I68-H69-V70IN439KD614G, Hits: 1.
I68MH69PV70IN439KD614GD1146Y, Hits: 1.
LSFL18FH49YN440KK444ND614G, Hits: 1.
Y144FA222VD614G, Hits: 1.
D614GT223IS940F, Hits: 1.
V551ID614G, Hits: 1.
F186SD614GD839Y, Hits: 1.
S221I, Hits: 1.
L176F, Hits: 1.
Q14HD614G, Hits: 1.
A879T, Hits: 1.
I68-H69-V70-S71-G72-T73-N74-G75-T76-D614G, Hits: 1.
S50LT95IF140-Y145-H146-K147-N148-Y453FE484KD614G, Hits: 1.
P26LDD614G, Hits: 1.
A222VA262SD614G, Hits: 1.
D614GE1072Q, Hits: 1.
V83FD614G, Hits: 1.
D614GA1078S, Hits: 1.
D574YD614G, Hits: 1.
Q52LD614G, Hits: 1.
D614GS708Y, Hits: 1.
Appendix B. P1, P2, P3, P4 PROOF

P1 PROOF:

Given that A and B are non-empty sets, then inequalities (B.1) and (B.2) are satisfied:

\[ 0 \leq |A \backslash B| \leq |A| \] (B.1)

\[ 0 \leq |B \backslash A| \leq |B| \] (B.2)

Or the size of the relative complement is less than or equal to the size of the set itself. Multiplying inequality (B.1) by \(|A|\) and (B.2) by \(|B|\), one gets:

\[ 0 \leq |A| |A \backslash B| \leq |A|^2 \] (B.3)

\[ 0 \leq |B| |B \backslash A| \leq |B|^2 \] (B.4)

From the sum of inequality (B.3) and inequality (B.4), one gets:

\[ 0 \leq |A| |A \backslash B| + |B| |B \backslash A| \leq |A|^2 + |B|^2 \] (B.5)

Since \(|A|^2 + |B|^2 > 0\), less than or equal operator in (B.5) won’t be affected if one divides (B.5) by \(|A|^2 + |B|^2\):

\[ 0 \leq \frac{|A| |A \backslash B| + |B| |B \backslash A|}{|A|^2 + |B|^2} \leq \frac{|A|^2 + |B|^2}{|A|^2 + |B|^2} = 1 \]

Or it was proved that: \(0 \leq d(A, B) \leq 1\).

P2 PROOF:

If sets A and B are equivalent, then relative complements of B in A and B in A are empty sets: \(A \backslash B = \emptyset\) and \(B \backslash A = \emptyset\), i.e. there are no elements that are contained in set B and vice versa. Since cardinality of an empty set equals 0 and A and B are non-empty sets: \(|A| > 0\) and \(|B| > 0\), we get that: \(|A \backslash B| = 0\), \(|B \backslash A| = 0\) and \(|A \backslash B| + |B \backslash A| = 0\). Eq. (1) nominator \(|A \\cup B|\) and \(|A \cap B| = 0\) divided by \(|A|^2 + |B|^2 > 0\) results in \(d(A, B) = 0\), given that \(A = B\).

Now I will prove property in opposite direction, or given that \(d(A, B) = 0\) then \(A = B\). Given that \(d(A, B) = 0\) and \(|A|^2 + |B|^2 > 0\), requires \(|A \backslash B|\) and \(|B \backslash A| = 0\) in Eq. (1). Since both sets are non-empty: \(|A| > 0\) and \(|B| > 0\), then \(|A \backslash B| + |B \backslash A| = 0\) iff \(|A \backslash B| = 0\) or \(|B \backslash A| = 0\). Set-theoretic differences equal empty set iff \(A = B = A \cap B\).

P3 PROOF:

In the case of disjoint sets A and B, there are no shared elements between sets \(A \cap B = \emptyset\). Because sets A and B do not intersect, relative complements equal sets: \(A \backslash B = A\) and \(B \backslash A = B\). Substituting \(|A \backslash B| = |A|\) and \(|B \backslash A| = |B|\) in Eq. (1) results in \(d(A, B) = |A| |B| / (|A|^2 + |B|^2) = |A| |B| / (|A|^2 + |B|^2) = 1\), so given that \(A = B = \emptyset\) then \(d(A, B) = 1\) was proved.

The opposite state: ‘if \(d(A, B) = 1\) then \(A \cap B = \emptyset\)’ and that will be proved in addition. If \(d(A, B) = 1\), then Eq. (1) can be transformed in: \(|A \backslash B| = |A| |B| / (|A|^2 + |B|^2) = 1\). Previous equation becomes true: for \(A \backslash B = A\) and \(B \backslash A = B\), while \(A \cap B = A = B = \emptyset\) if \(A\) and \(B\) are disjoint sets or \(A \cap B = \emptyset\).

P4 PROOF:

Given that \(d(B, A) = |B| |A| / (|A|^2 + |B|^2)\) by interchanging A with B in Eq. (1) and since \(d(B, A) = |B| |A| / (|A|^2 + |B|^2) = |A| |B| / (|A|^2 + |B|^2) = d(A, B)\), Eq. (1) symmetry is proved.

Appendix C. Condensed distance matrix for variants on Fig. 3 applying Eq. (1)
D. Stojanov

Davies, N.G., Jarvis, C.I., Edmunds, W.J., Jewell, N.P., Diaz-Ordaz, K., Keogh, R.H., Dice, L.R., 1945. Measures of the amount of ecologic association between species.

Choi, B., Choudhary, M.C., Regan, J., Sparks, J.A., Padera, R.F., Qiu, X., Solomon, I.H., Aydillo, T., Gonzalez-Reiche, A.S., Aslam, S., van de Guchte, A., Khan, Z., Obla, A., Davies, N.G., Abbott, S., Barnard, R.C., Jarvis, C.I., Kucharski, A.J., Munday, J.D., Avanzato, V.A., Matson, M.J., Seifert, S.N., Pryce, R., Williamson, B.N., Anzick, S.L., 2021b. Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. Nature 593 (7858), 270–274. https://doi.org/10.1038/s41586-021-03426-1.

References

Avanzato, V.A., Matson, M.J., Seifert, S.N., Pryce, R., Williamson, B.N., Anzick, S.L., Barbier, K., Judson, S.D., Fischer, E.R., Martens, C., Bowden, T.A., 2020. Case study: prolonged infectious SARS-CoV-2 shedding from an asymptomatic immunocompromised individual with cancer. Cell 183 (7), 1912. https://doi.org/10.1016/j.cell.2020.10.049.

Ashworth, J., Colquhoun, R., Connor, T.R., Faria, N.R., 2021. Establishment and lineage dynamics of the SARS-CoV-2 S epidemic in the UK. Science 371 (6511), 1603–1607. https://doi.org/10.1126/science.abf2946.

Bain, W.G., Jacobs, J., Nambulli, S., Parikh, U., Cillo, A., Staines, B., Hensley, M.K., 2020. The SARS-CoV-2 spike glycoprotein biosynthesis, structure, function, and antigenicity: implications for the design of spike-based vaccine immunogens. Front. Immunol. 11, 2593. https://doi.org/10.3389/fimmu.2020.576622.

Barbian, K., Judson, S.D., Fischer, E.R., Martens, C., Bowden, T.A., 2020. Case study: prolonged infectious SARS-CoV-2 shedding from an asymptomatic immunocompromised individual with cancer. Cell 183 (7), 1912. https://doi.org/10.1016/j.cell.2020.10.049.

Barbian, K., Judson, S.D., Fischer, E.R., Martens, C., Bowden, T.A., 2020. Case study: prolonged infectious SARS-CoV-2 shedding from an asymptomatic immunocompromised individual with cancer. Cell 183 (7), 1912. https://doi.org/10.1016/j.cell.2020.10.049.
Hoffmann, M., Kleine-Weber, H., Pohlmann, S., 2020. A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. Mol. Cell 78 (4), 779–784. https://doi.org/10.1016/j.molcel.2020.04.022.

Huang, Y., Yang, C., Xu, X.F., Xu, W., Liu, S.W., 2020. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol. Sin. 41 (9), 1141–1149. https://doi.org/10.1038/s41401-020-0485-4.

Jaccard, P., 1912. The distribution of the flora in the alpine zone. 1. New Phytol. 11 (2), 784. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

Lasek-Nesselquist, E., Pata, J., Schneider, E., George, K.S., 2021. A tale of three SARS-CoV-2 lineages: B.1.1.7, P.1, and B.1.351. Nature 593 (7857), 130–1310. https://doi.org/10.1038/s41586-021-02062-6.

Pen, R., Brown, J.C., Sanchez-David, R.Y., Braga, L., 2021. The furin cleavage site of SARS-CoV-2 is essential for infection of human lung cells. Mol. Cell 78 (4), 779–784. https://doi.org/10.1016/j.molcel.2020.04.022.

Rodrigues, C.H., Myung, Y., Pires, D.E., Ascher, D.B., 2019. mCSS-PP12: predicting the effects of mutations on protein-protein interactions. Nucleic Acids Res. 47 (W1), W338–W344. https://doi.org/10.1093/nar/gkz238.

Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A., Li, F., 2020. Cell entry mechanisms of SARS-CoV-2. Proc. Natl. Acad. Sci. 117 (21), 11727–11734. https://doi.org/10.1073/pnas.2003136117.

Sonnens, T.A., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Biol. Skar. 5, 1–34.

Starr, T.N., Greaney, A.J., Hilton, S.K., Ellis, D., Crawford, K.H., Dingens, A.S., Navarro, M.J., Bowen, J.E., Tortorici, M.A., Walls, A.C., King, N.F., 2020. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell 182 (5), 1295–1310. https://doi.org/10.1016/j.cell.2020.08.012.

Tanimoto, T.T., 1958. Elementary Mathematical Theory of Classification and Prediction. Volz, E., Hill, V., McCrone, J.T., Price, A., Jorgensen, D., O’Toole, Å., Southgate, J., Johnson, R., Jackson, B., Nascimento, F.F., Rey, S.M., 2021. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. Cell 184 (1), 64–75. https://doi.org/10.1016/j.cell.2020.11.026.

Wang, P., Nair, M.S., Liu, L., Iketani, S., Luo, Y., Guo, Y., Wang, M., Yu, J., Zhang, B., Kwong, P.D., Graham, B.S., 2021. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. Nature 593 (7857), 130–135. https://doi.org/10.1038/s41586-021-03998-2.

Ward Jr., J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58 (301), 236–244. https://doi.org/10.1080/01621459.1963.10500845.

Warwick, J., 2021. A model for pH coupling of the SARS-CoV-2 spike protein open/closed equilibrium. Brief. Bioinform. 22 (2), 1499–1507. https://doi.org/10.1093/bib/bbab056.

Wrap, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.L., Abiona, O., Graham, B., S., McElannan, J.S., 2020. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 367 (6483), 1260–1263. https://doi.org/10.1126/science.aab2507.