Computational Statistics
https://doi.org/10.1007/s00180-022-01291-8

ORIGINAL PAPER

Multiple partition Markov model for B.1.1.7, B.1.351, B.1.617.2, and P.1 variants of SARS-CoV 2 virus

Jesús Enrique García[1] · Verónica Andrea González-López[1] · Gustavo Henrique Tasca[2]

Received: 27 August 2021 / Accepted: 27 September 2022

Abstract

With tools originating from Markov processes, we investigate the similarities and differences between genomic sequences in *FASTA* format coming from four variants of the SARS-CoV 2 virus, B.1.1.7 (UK), B.1.351 (South Africa), B.1.617.2 (India), and P.1 (Brazil). We treat the virus' sequences as samples of finite memory Markov processes acting in $A = \{a, c, g, t\}$. We model each sequence, revealing some heterogeneity between sequences belonging to the same variant. We identified the five most representative sequences for each variant using a robust notion of classification, see Fernández et al. (Math Methods Appl Sci 43(13):7537–7549. https://doi.org/10.1002/mma.5705). Using a notion derived from a metric between processes, see García et al. (Appl Stoch Models Bus Ind 34(6):868–878. https://doi.org/10.1002/asmb.2346), we identify four groups, each group representing a variant. It is also detected, by this metric, global proximity between the variants B.1.351 and B.1.1.7. With the selected sequences, we assemble a multiple partition model, see Cordeiro et al. (Math Methods Appl Sci 43(13):7677–7691. https://doi.org/10.1002/mma.6079), revealing in which states of the state space the variants differ, concerning the mechanisms for choosing the next element in A. Through this model, we identify that the variants differ in their transition probabilities in eleven states out of a total of 256 states. For these eleven states, we reveal how the transition probabilities change from variant (group of variants) to variant (group of variants). In other words, we indicate precisely the stochastic reasons for the discrepancies.

Keywords Metric between Markov processes · Bayesian information criterion · Model selection · Model comparison

✉ Jesús Enrique García
jg@ime.unicamp.br

Verónica Andrea González-López
veronica@ime.unicamp.br

[1] Department of Statistics, University of Campinas, Sergio Buarque de Holanda, 651, Campinas, São Paulo CEP: 13083-859, Brazil

[2] Campinas, Brazil

Published online: 01 November 2022

1 Introduction

Recently developed stochastic process' tools are brought together in this paper to describe and model the stochastic behavior of four variants of the SARS-CoV 2 virus, these are: B.1.1.7 or Alpha variant (region of reference: UK), B.1.351, a member of Beta variant (region of reference: South Africa), B.1.617.2, a member of Delta variant (region of reference: India), and P.1, see Hirotsu and Omata (2021), a member of Gamma variant (region of reference: Brazil) (Hoffmann et al. 2021; Deng et al. 2021). The three variants B.1.1.7, B.1.351, and P.1 had coexisted and dominated since the end of 2020, taking over most of the infections until May of 2021, when they began to share the domain with a fourth variant B.1.617.2, considered responsible for the second wave of contamination in India. This last variant was first identified in October of 2020, but it gained impulse in February of 2021. A virus such as SARS-CoV 2 is prone to mutations, a wide range of variants classified in https://cov-lineages.org can be found, but some of these mutations make the virus more efficient, as is the case of the four variants mentioned here. In this paper, we restrict ourselves to analyzing four variants: B.1.1.7, B.1.351, B.1.617.2, and P.1, since these are considered *Variants of Concern* according to the US government SARS-CoV-2 Interagency Group (SIG), see for instance: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html. A variant in this classification implicates an increase in disease severity, a rise in transmissibility, misdiagnosis (confused symptoms), and potential decreased immunity from vaccination or previous disease.

As far as we know, existing vaccines tend to protect the population as they are effective in neutralizing the most widespread variants, but what has become hard is universal access to vaccines. Countries that produce vaccines for SARS-CoV 2 have the opportunity to adjust the vaccination process that their inhabitants have to receive to promote an increase in immunological capacity and, thus, be able to face variants of the SARS-CoV 2 virus. So far, other countries have not even managed to offer vaccines to their entire populations, which allows the virus to continue to evolve, adapting and defeating the human immune system, see Nonaka et al. (2021). This situation leads to the need for constant verification of the effectiveness of the existing vaccines in the new variants. It shows that the lack of equal access to health can compromise global social and economic developments since there is a network of exchange between different continents and countries. Then, all of humanity is affected by the evolution of the SARS-CoV 2 virus, despite the already existence of vaccines. In the literature can be found evidence showing how the rapid evolution of these variants in just one country affects entire continents. For example, what is placed in Taylor (2021) is that despite the fact that countries such as Uruguay and Chile had been controlling the pandemic until early 2021 when it got out of control in Brazil due to the prevalence of infections by P.1 variant, there has been a deterioration of the pandemic's control in Uruguay, Chile, and other Latin American countries. In other words, the global effort to maintain an adequate rate of immunization worldwide could be the only alternative to the current situation.

Since variants are known to be identified by genetic analysis, the question arises: why find models for them? Statistical models allow descriptions and representations of phenomena. Models can be used to compact information to make it reach repositories with high speed and quality (Christley et al. 2009). We see in the modeling process a solution for understanding the phenomenon that passes through a representation that allows its recovery with fidelity.

Models allow comparisons, which leads us to have a stochastic structure to understand and interpret the differences between variants, then this finding will have space in our paper.

We propose to consider complete genetic sequences from the four variants in *FASTA* format. Such sequences will be considered samples from discrete Markovian processes acting on the discrete genetic alphabet $\{a, c, g, t\}$, where a means adenine, c cytosine, g guanine, and t thymine. In this framework, we seek to verify which models are more and less compatible with the sequences' behavior of each variant, including investigating whether there is a considerable difference between one variant to another. We approach models such as those introduced in García and González-López (2017), García et al. (2020a). We also seek to represent the distance between the variants, which we will carry out using concepts such as those developed in García and González-López (2017), García et al. (2018). Finally, we will explore the aspects that the variants have in common or not, under the stochastic representation also proposed in this article and based on models such as those introduced in Cordeiro et al. (2020). Under this perspective, we seek to discriminate the variants through transition probabilities and states, of the state space of the Markovian processes, for which they are discrepant.

Precursors to the idea of understanding the genetic behavior as stochastic processes can be found in Buhlmann and Wyner (1999). In the current paper, we do not obtain a tree structure like the one obtained in Buhlmann and Wyner (1999), since the models used here are generalizations of *Variable Length Markov Chains* (Rissanen 1983; Csiszár and Talata 2006) for which we are still looking for a coherent graphic representation. It is necessary to observe that the estimation of a Variable Length Markov Chain can be consistently obtained through the Bayesian Information Criterion, as shown in Csiszár and Talata (2006). Moreover, in our case, the Bayesian Information Criterion produces a consistent estimation.

This paper is organized as follows. Section 2 is intended to introduce the theoretical framework of stochastic tools. We introduce and analyze the properties of a metric to compare samples from stochastic processes. We introduce and analyze the properties of a classifier of samples coming from stochastic processes. We introduce a model specifically designed to extract similarities and dissimilarities between samples from stochastic processes. In this section, we also detail what is expected to be achieved with such tools. Section 3.1 presents the data structure and some results which allow comparison. Sections 3.2 and 3.3 show the models by variant, and the joint model for the four variants. Section 4 presents the conclusions and findings of this paper.

2 Theoretical background

This section introduces three tools developed within the framework of stochastic processes. The first tool is a metric between samples from stochastic processes, based on García et al. (2018). Second a classifier of samples coming from stochastic processes, based on Fernández et al. (2020), and third, a model (Cordeiro et al. 2020) from which we will extract information about stochastic aspects that the four variants share and which the variants do not share.

The first challenge is to create a database that allows a fair comparison between the variants. We introduce stochastic tools to make a selection of sequences of each variant and later comparing them. This selection process becomes relevant because we must achieve a certain homogeneity in the set of sequences used to develop the models.

We introduce the theoretical context used for the investigation of the genetic structures of the virus. We consider the *FASTA* format, under such a format, a genetic sequence is represented by the concatenation of elements of the alphabet $A = \{a, c, g, t\}$. Each genetic sequence is identified as a sample coming from a stochastic process, which brings us to the context of stochastic processes on a discrete alphabet.

Let $(Z_t)_{t\geq 1}$ be a discrete time Markov chain of order o (finite) on a finite alphabet A. We denote by v_k^m the string $v_k v_{k+1} \dots v_m$, with $v_i \in A, i = k, \cdots, m$. Then, the state space of $(Z_t)_{t\geq 1}$ is A^o. Denote the transition probabilities over the state space A^o, as

$$P(v|s) = \text{Prob } (Z_t = v|Z_{t-o}^{t-1} = s), \quad v \in A, \quad s \in A^o. \tag{1}$$

The state space of a process can have a specific organization, in which there is a gap in the dependence from the past. For example, with the probabilities at time t depending on the values at times $t - o, ..., t - 1$, plus the value at time $t - G$ for a value $G : G > o$, as is the case introduced in García et al. (2020a). In that case, the transition probabilities are given by Eq. (2) over the state space $A \times A^o$,

$$P(v|zs) = \text{Prob } (Z_t = v|Z_{t-G} = z...Z_{t-o}^{t-1} = s), \quad v \in A \ \ z \in A, s \in A^o. \tag{2}$$

In the latter case, there is a part of the past, represented by '...' in Eq. (2) that is considered irrelevant. Given the diversity of possible structures, we propose the following notation to embrace the notions (1), (2), and others that can be incorporated. Let $(Z_t)_{t\geq 1}$ be a discrete time Markov chain on a finite alphabet A with state $\mathcal{S} = \Lambda(A)$, where $\Lambda(\cdot)$ denotes a function representing the memory structure, for example (i) $\Lambda(A) = A^o$, in Eq. (1), (ii) $\Lambda(A) = A \times A^o$, in Eq. (2). The transition probabilities are given by

$$P(v|s) = \text{Prob } (Z_t = v|\Omega(Z_{t-(\dots)}^{t-1}) = s), \quad v \in A, \quad s \in \Lambda(A), \tag{3}$$

where $\Omega(Z_{t-(\dots)}^{t-1})$ is the lapse at the past in which is observed the state s. For instance, (i) $\Omega(Z_{t-(\dots)}^{t-1}) = Z_{t-o}^{t-1}$ in Eq. (1), (ii) $\Omega(Z_{t-(\dots)}^{t-1}) = Z_{t-G}...Z_{t-o}^{t-1}$ in Eq. (2).

In practice, the transition probabilities are estimated from a sample, so consider a sample z_1^n of the process $(Z_t)_{t\geq 1}$, and in general terms denote by $N_n(s)$ all the occurrences of $s \in \Lambda(A)$ in the sample and denote by $N_n(s,v)$ all the occurrences of $s \in \Lambda(A)$ followed by $v \in A$ in the sample. Then, the estimator of $P(v|s)$ is given by

$$\hat{P}(v|s) = \frac{N_n(s,v)}{N_n(s)}, \quad v \in A, s \in \Lambda(A). \tag{4}$$

Note that following the situations given by Eqs. (1)–(2) we have, respectively: (i) $N_n(s) = |\{t : o < t \leq n, z_{t-o}^{t-1} = s\}|$ and $N_n(s,v) = |\{t : o < t \leq n, z_{t-o}^{t-1} = s, z_t = v\}|$, $s \in A^o,\ v \in A$; (ii) $N_n((z,s)) = |\{t : G < t \leq n, z_{t-G} = z, z_{t-o}^{t-1} = s\}|$ and $N_n((z,s),v) = |\{t : G < t \leq n, z_{t-G} = z, z_{t-o}^{t-1} = s, z_t = v\}|, (z,s) \in A \times A^o,\ v \in A$.

It is necessary to note that Eq. (1) could be enough, also in situations like the one given by Eq. (2). It is only necessary to increase the value of o to reach G. The aspect that needs to be considered is that as the memory increases, the sample size, n, must also be increased to obtain a consistent estimator of $P(v|s)$. From this perspective, it seems coherent to look for alternative models, such as the one offered by Eq. (2).

Next, our objective is to show that it is possible to measure the proximity between two samples from stochastic processes using the estimates of the transition probabilities introduced in Eq. (4).

2.1 Selection strategy

We formally present the equations used to quantify the proximity between samples from stochastic processes.

Definition 1 Consider two Markov chains $(Z_{1,t})$ and $(Z_{2,t})$ over a finite alphabet A, with state space $\Lambda(A)$ folowing Eq. (3). Consider two independent samples $z_{1,1}^{n_1}$, $z_{2,1}^{n_2}$ coming respectively from the two chains.

i. For $s \in \Lambda(A)$,

$$d_s(z_{1,1}^{n_1}, z_{2,1}^{n_2}) = \frac{\alpha}{(|A|-1)\ln(n_1+n_2)} \sum_{v \in A} \left\{ \sum_{l=1,2} N_{n_l}(s,v) \ln\left(\frac{N_{n_l}(s,v)}{N_{n_l}(s)}\right) - N_{n_1+n_2}(s,v) \ln\left(\frac{N_{n_1+n_2}(s,v)}{N_{n_1+n_2}(s)}\right) \right\},$$

$$\text{dmax}\ (z_{1,1}^{n_1}, z_{2,1}^{n_2}) = \max_{s \in \Lambda(A)} \{d_s(z_{1,1}^{n_1}, z_{2,1}^{n_2})\},$$

ii.

with $N_{n_1+n_2}(s,v) = N_{n_1}(s,v) + N_{n_2}(s,v)$, $N_{n_1+n_2}(s) = N_{n_1}(s) + N_{n_2}(s)$, where N_{n_1} and N_{n_2} are given as usual, computed from the samples $z_{1,1}^{n_1}$ and $z_{2,1}^{n_2}$ respectively, and α is a real and positive value.

This notion is introduced in García et al. (2018). The motivation was to compare the processes (through their samples) and through the Bayesian Information Criterion (BIC) (Schwarz 1978).

Under the assumptions of Definition 1, in García et al. (2018), the expressions of the BIC are compared in two situations, (a) vs. (b), the higher BIC points to the best model, since we assume the BIC expression given by Schwarz (1978). If P is related to the process $(Z_{1,t})$ and Q related to process $(Z_{2,t})$, in the situation (a), the BIC is calculated under the assumption that the processes share the same law concerning state $s, P(\cdot|s) = Q(\cdot|s)$, resulting on the following expression for the BIC criterion,

$$\sum_{v\in A, r\in\Lambda(A)\setminus\{s\}} \left\{ N_{n_1}(r,v)\ln\left(\frac{N_{n_1}(r,v)}{N_{n_1}(r)}\right) + N_{n_2}(r,v)\ln\left(\frac{N_{n_2}(r,v)}{N_{n_2}(r)}\right)\right\}$$
$$+\sum_{v\in A} N_{n_1+n_2}(s,v)\ln\left(\frac{N_{n_1+n_2}(s,v)}{N_{n_1+n_2}(s)}\right) - \frac{(|A|-1)(2|\Lambda(A)|-1)}{\alpha}\ln(n_1+n_2).$$

This means that instead of two sets of probabilities to be estimated, for s, $\{P(v|s)\}_{v\in A}$ and $\{Q(v|s)\}_{v\in A}$, for that state s, only one set must be estimated, which allows us to reduce the total number of probabilities to be estimated from $2(|A|-1)|\Lambda(A)|$ to $(|A|-1)(2|\Lambda(A)|-1)$, in the whole state space.

On the other hand, in the situation (b) the BIC is calculated under the assumption where the processes do not share the same law concerning state s, resulting on the following expression for the BIC criterion,

$$\sum_{v\in A, r\in\Lambda(A)} \left\{ N_{n_1}(r,v)\ln\left(\frac{N_{n_1}(r,v)}{N_{n_1}(r)}\right) + N_{n_2}(r,v)\ln\left(\frac{N_{n_2}(r,v)}{N_{n_2}(r)}\right)\right\} - \frac{2(|A|-1)|\Lambda(A)|}{\alpha}\ln(n_1+n_2).$$

By the comparison between (a) and (b) arises the notion given by Definition 1. $P(\cdot|s) = Q(\cdot|s)$, if and only if, d_s goes to 0, almost surely, when $\min(n_1,n_2)\to\infty$. And, if there is $v\in A : P(v|s)\neq Q(v|s), d_s$ goes to ∞, almost surely, when $\min(n_1,n_2)\to\infty$. Also, these properties are verified by *dmax*. Then, both notions of Definition 1, i. and ii., are asymptotically consistent to detect if the laws P and Q are the same or not. Of course, the notion *dmax* is a global notion that allows a general comparison between two samples, not only concerning any specific state. We can also see from García et al. (2018) that Definition 1-i. is a metric in the mathematical sense of the term. This is, it satisfies the three properties of a metric. (1) It is greater or equal to zero, being null when the empirical estimation of the transition probabilities coming from the two samples are identical, (2) it is symmetrical, (3) it verifies the triangular inequality, for the proof see Theorem 1 in García et al. (2018). (1) and (2) follow from the log-sum inequality, (3), the triangular inequality results from the relationship between Definition 1-i. and the relative entropy between the empirical laws involved in the construction of Definition 1-i. The asymptotic behavior of d_s

(when the samples' size grows enough) is properly described and proved in Theorem 3 of García et al. (2018). It is a consequence of Lemmas 6.2 and 6.3 provided in the work of Csiszár and Talata (2006).

For each variant, we have a set of sequences. Each set includes a wide variety of sequences' behavior, as we will see later. Then, to build a balanced base of genetic sequences per variant, we introduce a tool that allows us to select a balanced set of genomic sequences for each variant. Fernández et al. (2020) proposes for this purpose the use of the following notion.

Definition 2 Given a finite collection $\{z_{j,1}^{n_j}\}_{j=1}^{m}$ of samples from the processes $\{(Z_{j,t})\}_{j=1}^{m}$ with probabilities $\{P^j(v|s), v \in A, s \in \Lambda(A)\}_{j=1}^{m}$, P^j given by Eq. (3), $\forall j \in \{1, \dots, m\}$. For a fixed $i \in \{1, 2, ..., m\}$ define

$$U(z_{i,1}^{n_i}) = \text{median } \{\text{dmax } (z_{i,1}^{n_i}, z_{j,1}^{n_j}) \, : j \neq i, 1 \leq j \leq m\}. \tag{5}$$

Where, given a sequence $\{w_j\}_{j=1}^{l}$, median $\{w_j, 1 \leq j \leq l\} = w_{(k+1)}$ if $l = 2k + 1$ and median $\{w_j, 1 \leq j \leq l\} = \frac{w_{(k)}+w_{(k+1)}}{2}$ if $l = 2k$, for k an integer and $w_{(j)}$ denoting the jth order statistic of the collection $\{w_j\}_{j=1}^{l}$.

The asymptotic behavior of the index U (Definition 2) is investigated in Fernández et al. (2020), see Theorem 1 and Corollary 1. Due to the properties deduced for d_s and due to the consistency of *dmax*, the index U, related to each sample, converges almost surely to zero when the sample size goes to infinity if and only if, that sample is commanded by a law shared by the majority of the samples in the set of samples. Also, almost surely, U goes to infinity, when the sample size goes to infinity, if and only if, that sample has a law not shared by the majority of the samples in the set of samples.

Formally, in Fernández et al. (2020) it is shown that if we define $J_i = \{j \, : \, 1 \leq j \leq m, P^j = P^i\}$, that is the collection of processes sharing the law of process i, when $\min\{n_1, \cdots, n_m\} \to \infty$

$$U(z_{i,1}^{n_i}) \to 0 \;\Leftrightarrow\; |J_i| > \left\lceil \frac{m}{2} \right\rceil \tag{6}$$

and

$$U(z_{i,1}^{n_i}) \to \infty \;\Leftrightarrow\; |J_i| \leq \left\lceil \frac{m}{2} \right\rceil. \tag{7}$$

Where $\lceil r \rceil$ represents the smallest integer which is also larger than r and $|J_i|$ is the cardinal of the set J_i, for each i, $1 \leq i \leq m$.

Equation (6) reflects that the law of sequence i is shared by at least 50% of the samples if and only if the indicator of sample i, $U(z_{i,1}^{n_i})$ is arbitrarily small when the samples sizes are large enough. In counterpoint (see Eq. (7)), the law of sequence

i is not shared by 50% of the samples, if and only if the indicator $U(z_{i,1}^{n_i})$ acquires arbitrarily large values.

As discussed in Fernández et al. (2020), U, can be used to identify which sequences best represent the entire sequences' set. We will order the sequences according to the U values. This strategy will indicate the more and the less representative sequences in the whole sequences' set, under the assumption given by the next statement.

Assumption 2.1 Given a finite collection of processes $\{(Z_{j,t})\}_{j=1}^{m}$ with probabilities $\{P^j\}_{j=1}^{m}$, over the finite alphabet A, with state space $\Lambda(A)$; there is a law P^{j^*}, for some $1 \leq j^* \leq m$, such that, $|J_{j^*}| = |\{i{:}1 \leq i \leq m, P^i = P^{j^*}\}| > |J_j| = |\{i{:}1 \leq i \leq m, P^i = P^j\}|, \forall j \neq j^*, j \in \{1, \dots, m\}$. The law P^{j^*} is the majority law in $\{(Z_{j,t})\}_{j=1}^{m}$.

Given the set of samples $\{z_{j,1}^{n_j}\}_{j=1}^{m}$ coming from $\{(Z_{j,t})\}_{j=1}^{m}$ and under the Assumption 2.1, (1) for each $j \in \{1, 2, \cdots, m\}$ compute $U(z_{j,1}^{n_j})$, (2) identify $u_{(i)}$ the ith order statistic of $\{U(z_{j,1}^{n_j}), 1 \leq j \leq m\}$, for $i = 1, \dots, m$, (3) denote by $z_{(i),1}^{n_{(i)}}$ the sample related to $u_{(i)}$, for $i = 1, \cdots, m$. Finally $\{z_{(j),1}^{n_{(j)}}\}_{j=1}^{m}$ is the ordered set according to the values of U.

According to Theorem 2 of Fernández et al. (2020) and under the Assumption 2.1, the ordering reported here points to the samples that best represent the complete set, samples closest to all the others. Sample $z_{(1),1}^{n_{(1)}}$ is the most representative, sample $z_{(2),1}^{n_{(2)}}$ is the second most representative, and so on.

The notion U will be used to select sequences for each variant. We understand that this process is relevant to guarantee a certain degree of homogeneity per variant.

Section 2.2, presents the model that will allow a global comparison between the variants.

We propose to consider a model that identifies a partition in the space made up of the four variants and a state-space compatible with all the four variants. The partition will gather in each of its parts, states of the state space and an index associated with the state that indicates to which variant that state is related. The elements in each part share their transition probabilities to any element of the genomic alphabet $A = \{a, c, g, t\}$. The identification of this partition will allow the use of different states and different variants to estimate the same transition probabilities, indicating which states of the state space the variants behave similarly and for which they do not. In other words, it is through these parts that we reveal the similarities and the discrepancies between the variants. And at the same time, we can report the computed transition probabilities more accurately as we use more samples for their estimation.

We extract from this model information showing what these variants have in common and what they do not share. The perspective of this model is naturally stochastic, and all the elements we use are related to state space, its organization, and transition probabilities.

2.2 Multiple partition Markov model

Given a collection of m independent processes $\mathcal{D} = \{(Z_{j,t})\}_{j=1}^{m}$, all Markov chains over a finite and discrete alphabet A, with state space $\Lambda(A)$, and transition probabilities $\{P^j(v|s), v \in A, s \in \Lambda(A)\}_{j=1}^{m}$, P^j given by Eq. (3), $\forall j \in \{1, \dots, m\}$. In the next definition, we give the structure over the state space $\Lambda(A)$ that needs to be identified and that determines the model to detect what the m processes in $\mathcal{D}$ have in common and what they do not have in common.

Definition 3 Given a collection of m independent processes $\mathcal{D} = \{(Z_{j,t})\}_{j=1}^{m}$, all Markov chains over a finite and discrete alphabet A, with state space $\Lambda(A)$, with probabilities $\{P^j(v|s), v \in A, s \in \Lambda(A)\}_{j=1}^{m}$, P^j given by Eq. (3), $\forall j \in \{1, \dots, m\}$.

1. $(i,\ s)$ and $(j,\ r)$, such that $(i,s),(j,r) \in \{1, \dots, m\} \times \Lambda(A)$ are equivalents if $P^i(\cdot|s) = P^j(\cdot|r)$.
2. $\mathcal{D}$ has a partition $\mathbb{P} = \{\mathcal{P}_1, \dots, \mathcal{P}_{|\mathbb{P}|}\}$ on $\{1, \dots, m\} \times \Lambda(A)$, if this is the one following the equivalence of 1. on the space $\{1, \dots, m\} \times \Lambda(A)$.

Note that item 2. of Definition 3 means that the parts $\mathcal{P}_j$, $1 \le j \le |\mathbb{P}|$ of $\mathbb{P}$ contain elements following 1. and there is no two parts $\mathcal{P}_j, \mathcal{P}_i$, $i \neq j$ sharing the transition probabilities.

In a model like this, proposed in Cordeiro et al. (2020), we sought to identify the states where the performance of two processes is the same (say $P^i(\cdot|s) = P^j(\cdot|r)$). That is done previously identifying the partition $\mathbb{P}$. Note that we are then using several states and processes to infer the same parameter (in this case, the parameters are probabilities), then for instance $\forall (i,s) \in \mathcal{P}$ some part of $\mathbb{P}$, $P(\cdot|\mathcal{P}) = P^i(\cdot|s)$.

Given a set of samples $\left\{ z_{j,1}^{n_j} \right\}_{j=1}^{m}$ of the set of processes $\mathcal{D} = \{(Z_{j,t})\}_{j=1}^{m}$ ($z_{j,1}^{n_j}$ is a sample of the process $(Z_{j,t})$), $\mathbb{P}$ is estimated by using the Bayesian Information Criterion (BIC). The BIC criterion makes it possible to consistently estimate the partition, since Theorem 3 of Cordeiro et al. (2020) shows that the BIC criterion allows to retrieve the partition $\mathbb{P}$, eventually almost surely, when the sample sizes $n_1, n_2, \dots, n_m$ go to infinity.

In practice, the estimation of $\mathbb{P}$ is obtained using a BIC-based notion introduced in Cordeiro et al. (2020), we introduce the notion in the following definition.

Definition 4 Under the assumptions of Definition 3, let $\left\{ z_{l,1}^{n_l} \right\}_{l=1}^{m}$ be samples of the collection $\mathcal{D}$, for $(i,s),(j,r) \in \{1, \cdots, m\} \times \Lambda(A)$, set

$$d\Big((i,s),(j,r)\Big) = \frac{2}{(|A|-1)\ln(\sum_{l=1}^{m} n_l)} \sum_{v \in A} \left\{ \sum_{k=i,j} \left\{ N((k,s_k),v) \ln\left(\frac{N((k,s_k),v)}{N(k,s_k)} \right) \right\} - N(\{(i,s),(j,r)\},v) \ln\left(\frac{N(\{(i,s),(j,r)\},v)}{N(\{(i,s),(j,r)\})} \right) \right\},$$

where $N(\{(i,s),(j,r)\}) = N_{n_i}(s) + N_{n_j}(r),\ N(\{(i,s),(j,r)\},v) = N_{n_i}(s,v) + N_{n_j}(r,v), s_i = s$ and $s_j = r$.

The notion given by Definition 4 is a metric in $\{1,\cdots,m\} \times \Lambda(A)$ and can be used in algorithms, to identify the partition $\mathbb{P}$ established by Definition 3-2. The condition of metric is formally established in Theorem 2 of Cordeiro et al. (2020). The consistency of the metric is proved in Theorem 1 of Cordeiro et al. (2020), a property that support its use (for sample sizes large enough) to identify the composition of the partition $\mathbb{P}$.

Once $\mathbb{P}$ is estimated by $\hat{\mathbb{P}}$, the set of transition probabilities $\{P(\cdot|\mathcal{P}),\ \mathcal{P} \in \mathbb{P}\}$ is estimated by $\{\hat{P}(\cdot|\hat{\mathcal{P}}),\ \hat{\mathcal{P}} \in \hat{\mathbb{P}}\}$, with

$$\hat{P}(v|\hat{\mathcal{P}}) = \frac{\sum_{(i,s)\in\hat{\mathcal{P}}} N_{n_i}(s,v)}{\sum_{(i,s)\in\hat{\mathcal{P}}} N_{n_i}(s)}, \quad \hat{\mathcal{P}} \in \hat{\mathbb{P}},\ v \in A. \tag{8}$$

That is, different elements of $\Lambda(A)$ and different samples are used to estimate the same transition probability.

Since we will consider the genomic sequences over $A = \{a, c, g, t\}$ as being samples coming from stochastic processes. The stochastic tools and models introduced in this section allow (1) to establish a state-space, $\Lambda(A)$, compatible with all the selected genomic sequences, (2) to identify variant by variant the most representative sequence (minimun value of U) and the least representative sequence (maximum value of U), giving a statistical meaning for such classification, (3) to establish a context (partition $\mathbb{P}$) that is used to identify the states for which the variants behave similarly and the states for which the variants disagree, also giving a stochastic meaning to the concept of similarity/dissimilarity.

The following section describes the database and its source in Sect. 3.1. In Sect. 3.2, we present marginal models for each sequence, a comparative analysis between the sequences, and the selection of the most representative sequences per variant. In Sect. 3.3 we present the joint model that aims to identify the commonalities and the differences between the variants.

3 Data and models

We consider 15 genetic sequences per variant, collected on dates close to each other, in order to capture the situation in that period of time. Next, we investigate possible patterns for each of these 15 sequences by variant. We select five sequences per variant that allow a more coherent comparison between the variants. Finally, we fit a model for the four variants revealing both similarities and divergences.

Table 1 Collection of sequences of SARS-CoV 2, obtained from GISAID source

B.1.1.7			B.1.351		
Accession ID	Collection date	Length	Accession ID	Collection date	Length
EPI_ISL_2753544	2021-01-28	29,846	EPI_ISL_2360687	2021-01-07	29,847
EPI_ISL_2753547	2021-01-30	29,805	EPI_ISL_2360709	2021-01-05	29,801
EPI_ISL_2753549	2021-01-30	29,805	EPI_ISL_2360747	2021-01-05	29,847
EPI_ISL_2753550	2021-01-25	29,829	EPI_ISL_2375982	2021-01-03	29,845
EPI_ISL_2753551	2021-01-25	29,806	EPI_ISL_2375983	2021-01-06	29,846
EPI_ISL_2753554	2021-01-30	29,805	EPI_ISL_2582710	2021-01-16	29,460
EPI_ISL_2753555	2021-01-30	29,868	EPI_ISL_2582720	2021-01-16	29,751
EPI_ISL_2753559	2021-01-25	29,814	EPI_ISL_2621127	2021-01-07	29,846
EPI_ISL_2753563	2021-01-28	29,835	EPI_ISL_2662514	2021-01-11	29,602
EPI_ISL_2753565	2021-01-30	29,807	EPI_ISL_2662518	2021-01-11	29,751
EPI_ISL_2753569	2021-01-28	29,805	EPI_ISL_2662528	2021-01-07	29,751
EPI_ISL_2753570	2021-01-28	29,830	EPI_ISL_2662529	2021-01-06	29,751
EPI_ISL_2753572	2021-01-25	29,814	EPI_ISL_2662531	2021-01-06	29,751
EPI_ISL_2753574	2021-01-25	29,812	EPI_ISL_2662533	2021-01-06	29,751
EPI_ISL_2753613	2021-01-05	29,817	EPI_ISL_2662550	2021-01-27	29,751
B.1.617.2			P.1		
Accession ID	Collection date	Length	Accession ID	Collection date	Length
EPI_ISL_1662291	2021-03-14	29,887	EPI_ISL_2777405	2021-01-15	29,593
EPI_ISL_1663304	2021-03-15	29,782	EPI_ISL_2777406	2021-01-15	29,730
EPI_ISL_1663312	2021-03-15	29,845	EPI_ISL_2777414	2021-01-27	29,593
EPI_ISL_1663376	2021-03-01	29,849	EPI_ISL_2777416	2021-01-29	29,728
EPI_ISL_1663501	2021-03-28	29,782	EPI_ISL_2777418	2021-01-19	29,593
EPI_ISL_1663502	2021-03-28	29,855	EPI_ISL_2777454	2021-01-12	29,784
EPI_ISL_1663507	2021-03-16	29,783	EPI_ISL_2777470	2021-01-13	29,635
EPI_ISL_1663522	2021-03-22	29,901	EPI_ISL_2777471	2021-01-14	29,743
EPI_ISL_1663541	2021-03-22	29,840	EPI_ISL_2777477	2021-01-25	29,758
EPI_ISL_1663557	2021-03-25	29,820	EPI_ISL_2777478	2021-01-25	29,642
EPI_ISL_1663563	2021-03-25	29,797	EPI_ISL_2777479	2021-01-26	29,735
EPI_ISL_1663564	2021-03-29	29,825	EPI_ISL_2777480	2021-01-26	29,784
EPI_ISL_1704234	2021-03-08	29,800	EPI_ISL_2777507	2021-01-28	29,777
EPI_ISL_1704618	2021-03-11	29,800	EPI_ISL_2777509	2021-01-31	29,774
EPI_ISL_1704629	2021-03-15	29,793	EPI_ISL_2777552	2021-01-13	29,784

3.1 Data

The database consists of a collection of genetic sequences in *FASTA* format. For that reason, the alphabet that is considered is the genomic one, $A = \{a, c, g, t\}$. The complete genome sequences of SARS-CoV 2 used in this paper can be found in GISAID source (https://gisaid.org), the sequences are listed in Table 1. We identify the originating/submitting lab for each sequence, as informed and required by GISAID source, see Sect. 1. Table 1 records the *Accession ID* of each sequence, the collection data and the sample sizes (length). We see that all the sequences were collected at the beginning of 2021 and the sample sizes are greater than 29,500, that is, the number of elements concatenated and coming from $A = \{a, c, g, t\}$ is greater than 29,500, for each sequence. It should be noticed here that there is a limitation for the order of a Markov chain that can be achieved in the application, which depends on the size of the sample n, the constrain is $o < \lfloor \log_{|A|}(n) \rfloor - 1$, which in our case implies $o < 6$. Later we will discuss the choice of the order o that is shown to be reasonable in this database.

In the next section, we describe a first sequence-by-sequence inspection, which is carried out to identify the stochastic behavior of each one of them. Also, this first approach allows a comparison between the sequences.

3.2 Marginal models

For each sequence listed in Table 1, we fit models type as introduced by Definition 3-2. with $m = 1$. That is to say, that the partition in each case is composed of states that share their transition probabilities. The state-space $\Lambda(A)$ is given by structures like the one exposed in Eq. (2), i.e. $\Lambda(A) = A \times A^o$.

The genetic structures are organized in triplets, then suitable values for the order o are 3, 6, 9, etc, since o is the order responsible for the continuous memory of the process. We follow the heuristic rule in Markov processes modeling, which restrict the values of o to the sample size n and depending on the size's alphabet |*A*|, that is, $o < \lfloor \log_{|A|}(n) \rfloor - 1$. In the case of genetic samples (according to the sizes reported in Table 1) o would have to be less than 6, so the reasonable option is to use $o = 3$. The BIC criterion is then used to select the value of G, among the options $G = 4, 5, 6, 7, 8, 9, 10, 11, 12$. It makes sense to consider these scenarios, as they have been investigated in other papers, to represent genetic sequences of SARS-CoV 2, see for instance García et al. (2020a, b). A detailed discussion on the determination of o and G in models given by Definition 3-2. ($m = 1$) using the first records of genomic sequences of SARS CoV 2 is found in García et al. (2020a), in that paper, several values of o and G are explored through two selection criteria: BIC and Krichevsky-Trofimov (KT) criterion. The findings of García et al. (2020a) support the values indicated here.

Tables 2, 3, 4, and 5 show the results of the BIC values, by variant, B.1.1.7, B.1.351, B.1.617.2 and P.1, respectively. To the right of each sequence the 4 best models are listed (the higher the BIC, the more indicated the model). We highlight in bold, the most representative sequences. We explain later how they were identified.

Regarding the variant B.1.1.7 (UK) we see that the predilection is for the value $G = 9$, with a possible exception given by the sequence EPI_ISL_2753569 in which case $G = 9$ is pointed as the second-best option after $G = 12$. Anyway, this fact is following the literature, see García et al. (2020a, b), where $G = 9$ is pointed to describe the original sequence of SARS-CoV 2 virus, ID: MN908947—China (Wuhan), obtained from NCBI (https://www.ncbi.nlm.nih.gov/nuccore), among others (García et al. 2020b).

Table 2 By line, for each sequence of variant B.1.1.7, the four highest BIC values in the comparison between models, see Definition 3, $m = 1, \Lambda(A) = A \times A^o, o = 3$ and $G = 4, \dots, 12$ (see Eq. (2)). In bold letter the most representative sequences (see Table 6)

Sequence	1st Model (BIC)	2nd Model (BIC)	3rd Model (BIC)	4th Model (BIC)
EPI_ISL_2753544	– 39563.4	– 39566.7	– 39567.1	– 39567.4
	$G = 9$	$G = 4$	$G = 6$	$G = 10$
EPI_ISL_2753547	– 39507.6	– 39517.2	– 39525.8	– 39527.0
	$G = 9$	$G = 4$	$G = 10$	$G = 12$
EPI_ISL_2753549	– 39509.0	– 39524.1	– 39525.8	– 39527.4
	$G = 9$	$G = 5$	$G = 4$	$G = 10$
EPI_ISL_2753550	– 39536.9	– 39553.3	– 39554.9	– 39558.9
	$G = 9$	$G = 4$	$G = 10$	$G = 12$
EPI_ISL_2753551	– 39490.3	– 39510.0	– 39510.7	– 39512.4
	$G = 9$	$G = 6$	$G = 4$	$G = 12$
EPI_ISL_2753554	– 39373.5	– 39377.0	– 39384.3	– 39389.2
	$G = 9$	$G = 12$	$G = 10$	$G = 11$
EPI_ISL_2753555	– 39576.5	– 39586.4	– 39589.2	– 39593.3
	$G = 9$	$G = 12$	$G = 4$	$G = 10$
EPI_ISL_2753559	– 39527.5	– 39534.5	– 39539.7	– 39542.8
	$G = 9$	$G = 4$	$G = 10$	$G = 6$
EPI_ISL_2753563	– 39491.3	– 39509.4	– 39523.6	– 39527.2
	$G = 9$	$G = 12$	$G = 10$	$G = 11$
EPI_ISL_2753565	– 39235.9	– 39240.6	– 39255.8	– 39263.6
	$G = 9$	$G = 12$	$G = 10$	$G = 11$
EPI_ISL_2753569	– 39329.4	– 39331.7	– 39337.4	– 39338.7
	$G = 12$	$G = 9$	$G = 10$	$G = 11$
EPI_ISL_2753570	– 39543.1	– 39548.2	– 39553.1	– 39561.6
	$G = 9$	$G = 4$	$G = 10$	$G = 12$
EPI_ISL_2753572	– 39504.4	– 39527.9	– 39530.4	– 39532.9
	$G = 9$	$G = 12$	$G = 4$	$G = 6$
EPI_ISL_2753574	– 39519.1	– 39525.5	– 39538.6	– 39542.5
	$G = 9$	$G = 4$	$G = 6$	$G = 10$
EPI_ISL_2753613	– 39523.6	– 39535.0	– 39539.9	– 39543.7
	$G = 9$	$G = 6$	$G = 4$	$G = 10$

Concerning the variant B.1.351 (South Africa), we see that the two most indicated models point to $G = 9$ and $G = 12$, and these alternate across the sequences. With possible exceptions (EPI_ISL_2621127 and EPI_ISL_2662514), but even in those cases, between the four most indicated models are those with $G = 9$ and $G = 12$. When comparing Tables 3 and 2 we see that the first case exhibits greater diversity. It could point to structural changes in genomic sequences.

Variant B.1.617.2 (India) has a slightly more varied profile than the other two. The sequences admit 4 and 5 (among others) as ideal values of G, but all the

Table 3 By line, for each sequence of variant B.1.351, the four highest BIC values in the comparison between models, see Definition 3, $m = 1, \Lambda(A) = A \times A^o, o = 3$ and $G = 4, \ldots, 12$ (see Eq. (2))

Sequence	1st Model (BIC)	2nd Model (BIC)	3rd Model (BIC)	4th Model (BIC)
EPI_ISL_2360687	− 39146.7	− 39153.0	− 39165.0	− 39169.4
	$G = 9$	$G = 12$	$G = 11$	$G = 10$
EPI_ISL_2360709	− 39299.0	− 39313.6	− 39321.5	− 39324.2
	$G = 12$	$G = 9$	$G = 11$	$G = 10$
EPI_ISL_2360747	− 39162.8	− 39170.8	− 39185.4	− 39193.4
	$G = 12$	$G = 9$	$G = 10$	$G = 11$
EPI_ISL_2375982	− 39278.3	− 39288.1	− 39290.9	− 39298.1
	$G = 12$	$G = 9$	$G = 10$	$G = 6$
EPI_ISL_2375983	− 39291.2	− 39293.4	− 39295.1	− 39299.9
	$G = 12$	$G = 9$	$G = 10$	$G = 6$
EPI_ISL_2582710	− 38683.9	− 38706.2	− 38716.4	− 38720.8
	$G = 12$	$G = 9$	$G = 10$	$G = 11$
EPI_ISL_2582720	− 39204.2	− 39220.7	− 39233.0	− 39236.5
	$G = 12$	$G = 9$	$G = 10$	$G = 11$
EPI_ISL_2621127	− 39281.0	− 39281.7	− 39281.7	− 39283.0
	$G = 9$	$G = 10$	$G = 12$	$G = 6$
EPI_ISL_2662514	− 38734.4	− 38773.5	− 38776.7	− 38779.9
	$G = 12$	$G = 11$	$G = 10$	$G = 9$
EPI_ISL_2662518	− 39333.3	− 39344.7	− 39360.7	− 39361.7
	$G = 12$	$G = 9$	$G = 11$	$G = 10$
EPI_ISL_2662528	− 39364.9	− 39369.2	− 39383.0	− 39386.6
	$G = 12$	$G = 9$	$G = 10$	$G = 11$
EPI_ISL_2662529	− 39363.2	− 39364.9	− 39383.0	− 39387.4
	$G = 9$	$G = 12$	$G = 11$	$G = 10$
EPI_ISL_2662531	− 39366.2	− 39366.7	− 39372.5	− 39393.1
	$G = 9$	$G = 12$	$G = 10$	$G = 11$
EPI_ISL_2662533	− 39356.9	− 39366.1	− 39374.0	− 39394.5
	$G = 9$	$G = 12$	$G = 10$	$G = 11$
EPI_ISL_2662550	− 39360.7	− 39368.5	− 39388.1	− 39389.0
	$G = 9$	$G = 12$	$G = 11$	$G = 10$

In bold letter the most representative sequences (see Table 6)

sequences maintain the case $G = 9$ among the four best models, in the range considered from 4 to 12.

As seen in the variant B.1.617.2, the variant P.1 shows a greater diversity of ideal values of G, maintaining the value $G = 9$ among its four most indicated, except in the case of the sequence EPI_ISL_2777414.

As a consequence of the diversity observed between the sequences within each variant, the first situation that arises is defining a more homogeneous group of sequences per variant.

Table 4 By line, for each sequence of variant B.1.617.2, the four highest BIC values in the comparison between models, see Definition 3, $m = 1, \Lambda(A) = A \times A^o, o = 3$ and $G = 4, \ldots, 12$ (see Eq. (2))

Sequence	1st Model (BIC)	2nd Model (BIC)	3rd Model (BIC)	4th Model (BIC)
EPI_ISL_1662291	− 39611.7	− 39616.3	− 39622.9	− 39632.9
	$G = 9$	$G = 4$	$G = 12$	$G = 6$
EPI_ISL_1663304	− 39469.6	− 39472.0	− 39472.5	− 39472.8
	$G = 4$	$G = 9$	$G = 12$	$G = 5$
EPI_ISL_1663312	− 39492.2	− 39498.8	− 39499.9	− 39501.6
	$G = 9$	$G = 4$	$G = 12$	$G = 10$
EPI_ISL_1663376	− 39571.4	− 39574.0	− 39577.0	− 39580.0
	$G = 9$	$G = 4$	$G = 10$	$G = 12$
EPI_ISL_1663501	− 39486.3	− 39488.8	− 39489.5	− 39493.5
	$G = 4$	$G = 5$	$G = 12$	$G = 9$
EPI_ISL_1663502	− 39581.4	− 39581.5	− 39588.5	− 39590.2
	$G = 4$	$G = 5$	$G = 9$	$G = 12$
EPI_ISL_1663507	− 39478.1	− 39480.7	− 39499.1	− 39499.8
	$G = 4$	$G = 9$	$G = 10$	$G = 5$
EPI_ISL_1663522	− 39632.0	− 39636.7	− 39651.6	− 39653.5
	$G = 9$	$G = 4$	$G = 12$	$G = 6$
EPI_ISL_1663541	− 39558.6	− 39568.6	− 39572.3	− 39574.5
	$G = 4$	$G = 9$	$G = 12$	$G = 5$
EPI_ISL_1663557	− 39536.3	− 39542.7	− 39542.7	− 39546.7
	$G = 4$	$G = 5$	$G = 10$	$G = 9$
EPI_ISL_1663563	− 39500.0	− 39503.9	− 39511.6	− 39517.0
	$G = 4$	$G = 9$	$G = 5$	$G = 12$
EPI_ISL_1663564	− 39544.2	− 39546.1	− 39554.4	− 39558.0
	$G = 4$	$G = 9$	$G = 5$	$G = 10$
EPI_ISL_1704234	− 39439.1	− 39444.0	− 39462.7	− 39466.7
	$G = 12$	$G = 9$	$G = 10$	$G = 11$
EPI_ISL_1704618	− 39445.8	− 39450.3	− 39469.1	− 39470.3
	$G = 9$	$G = 12$	$G = 10$	$G = 11$
EPI_ISL_1704629	− 39453.2	− 39457.5	− 39471.0	− 39473.7
	$G = 9$	$G = 12$	$G = 10$	$G = 5$

In bold letter the most representative sequences (see Table 6)

As discussed in Sect. 2.1, we apply the classifier U with $m = 15$ per variant, see Definition 2, selecting the five best classified per variant. For such a process, we need to adopt a value G, we use $G = 9$ (and as earlier $o = 3$). That is, we have adopted as a common state space for the sequences, $\Lambda(A) = A \times A^o$, with $o = 3$ and $G = 9$. The classification of the sequences, by variant, is found in Table 6. On top is the most representative (with lower U value), on the bottom, is the least representative (with higher U value).

The U values reported in Table 6 indicate different amplitudes per variant. For variant B.1.1.7 (UK), the indicators belong to the interval [0.005809, 0.038354] with amplitude 0.032545. For variant B.1.351 (South Africa), the U values are in

Table 5 By line, for each sequence of variant P.1, the four highest BIC values in the comparison between models, see Definition 3, $m = 1, \Lambda(A) = A \times A^{o}, o = 3$ and $G = 4, \ldots, 12$ (see Eq. (2))

Sequence	1st Model (BIC)	2nd Model (BIC)	3rd Model (BIC)	4th Model (BIC)
EPI_ISL_2777405	− 39218.5	− 39224.9	− 39225.7	− 39235.6
	$G = 4$	$G = 12$	$G = 9$	$G = 5$
EPI_ISL_2777406	− 39411.1	− 39414.6	− 39419.8	− 39430.9
	$G = 9$	$G = 4$	$G = 12$	$G = 5$
EPI_ISL_2777414	− 39222.1	− 39224.8	− 39233.4	− 39239.5
	$G = 12$	$G = 8$	$G = 4$	$G = 5$
EPI_ISL_2777416	− 39405.6	− 39410.6	− 39415.3	− 39425.5
	$G = 9$	$G = 4$	$G = 12$	$G = 6$
EPI_ISL_2777418	− 39212.3	− 39221.1	− 39224.1	− 39228.8
	$G = 9$	$G = 4$	$G = 12$	$G = 5$
EPI_ISL_2777454	− 39478.7	− 39487.0	− 39497.6	− 39498.5
	$G = 4$	$G = 9$	$G = 5$	$G = 10$
EPI_ISL_2777470	− 39271.4	− 39274.5	− 39275.5	− 39285.6
	$G = 9$	$G = 12$	$G = 4$	$G = 10$
EPI_ISL_2777471	− 39427.4	− 39427.7	− 39440.1	− 39444.3
	$G = 4$	$G = 9$	$G = 12$	$G = 6$
EPI_ISL_2777477	− 39449.7	− 39451.0	− 39462.3	− 39469.6
	$G = 4$	$G = 9$	$G = 12$	$G = 6$
EPI_ISL_2777478	− 39281.5	− 39292.1	− 39293.4	− 39300.0
	$G = 9$	$G = 12$	$G = 4$	$G = 5$
EPI_ISL_2777479	− 39412.8	− 39423.0	− 39430.8	− 39434.0
	$G = 9$	$G = 4$	$G = 12$	$G = 5$
EPI_ISL_2777480	− 39482.5	− 39494.1	− 39494.1	− 39498.8
	$G = 9$	$G = 4$	$G = 12$	$G = 6$
EPI_ISL_2777507	− 39475.1	− 39481.6	− 39486.1	− 39492.0
	$G = 4$	$G = 12$	$G = 9$	$G = 5$
EPI_ISL_2777509	− 39474.5	− 39477.4	− 39481.2	− 39483.4
	$G = 9$	$G = 12$	$G = 4$	$G = 10$
EPI_ISL_2777552	− 39481.6	− 39489.9	− 39495.3	− 39497.9
	$G = 9$	$G = 4$	$G = 12$	$G = 5$

In bold letter the most representative sequences (see Table 6)

[0.011423, 0.07413] with amplitude 0.062707. The U values of variant B.1.617.2 (India), are in [0.005491, 0.079899] with amplitude 0.074408, and the indicators of variant P.1 (Brazil) are in [0.009976, 0.013368] with amplitude 0.003392. It shows that for variant P.1, the sequences can be considered closest to each other, while the sequence set of variant B.1.617.2 shows the greatest diversity.

Considering the five most representative sequences per variant, at the top of the lists in Table 6 and in bold in Tables 2, 3, 4, and 5, we see that all of them show,

Table 6 Accession ID and U values per variant, see Definition 2, $\Lambda(A) = A \times A^o$, with $o = 3$ and $G = 9$. From left to right from top to bottom, B.1.1.7, B.1.351, B.1.617.2 and P.1 variants

B.1.1.7 sequence	U	B.1.351 sequence	U	B.1.617.2 sequence	U	P.1 sequence	U
EPI_ISL_2753570	0.005809	EPI_ISL_2360709	0.011423	EPI_ISL_1663502	0.005491	EPI_ISL_2777471	0.009976
EPI_ISL_2753559	0.006287	EPI_ISL_2662528	0.011599	EPI_ISL_1663557	0.005907	EPI_ISL_2777454	0.009976
EPI_ISL_2753549	0.006939	EPI_ISL_2662531	0.011599	EPI_ISL_1663304	0.005908	EPI_ISL_2777480	0.009976
EPI_ISL_2753551	0.007512	EPI_ISL_2582720	0.011601	EPI_ISL_1704618	0.006278	EPI_ISL_2777552	0.009976
EPI_ISL_2753613	0.007749	EPI_ISL_2662529	0.011776	EPI_ISL_1663501	0.006278	EPI_ISL_2777479	0.009977
EPI_ISL_2753544	0.008258	EPI_ISL_2662533	0.012450	EPI_ISL_1663507	0.006278	EPI_ISL_2777507	0.009977
EPI_ISL_2753550	0.008264	EPI_ISL_2662550	0.013945	EPI_ISL_1663312	0.006438	EPI_ISL_2777509	0.009977
EPI_ISL_2753565	0.008327	EPI_ISL_2662518	0.013945	EPI_ISL_1663541	0.006691	EPI_ISL_2777406	0.009977
EPI_ISL_2753547	0.008556	EPI_ISL_2662514	0.014583	EPI_ISL_1704234	0.007106	EPI_ISL_2777416	0.009977
EPI_ISL_2753554	0.008706	EPI_ISL_2582710	0.019783	EPI_ISL_1663563	0.007338	EPI_ISL_2777477	0.009977
EPI_ISL_2753572	0.009383	EPI_ISL_2375982	0.060754	EPI_ISL_1663564	0.009255	EPI_ISL_2777478	0.013367
EPI_ISL_2753574	0.009383	EPI_ISL_2375983	0.065704	EPI_ISL_1704629	0.009256	EPI_ISL_2777470	0.013367
EPI_ISL_2753569	0.009708	EPI_ISL_2621127	0.065705	EPI_ISL_1663522	0.077592	EPI_ISL_2777405	0.013368
EPI_ISL_2753563	0.020767	EPI_ISL_2360747	0.073353	EPI_ISL_1662291	0.079895	EPI_ISL_2777418	0.013368
EPI_ISL_2753555	0.038354	EPI_ISL_2360687	0.074113	EPI_ISL_1663376	0.079899	EPI_ISL_2777414	0.013368

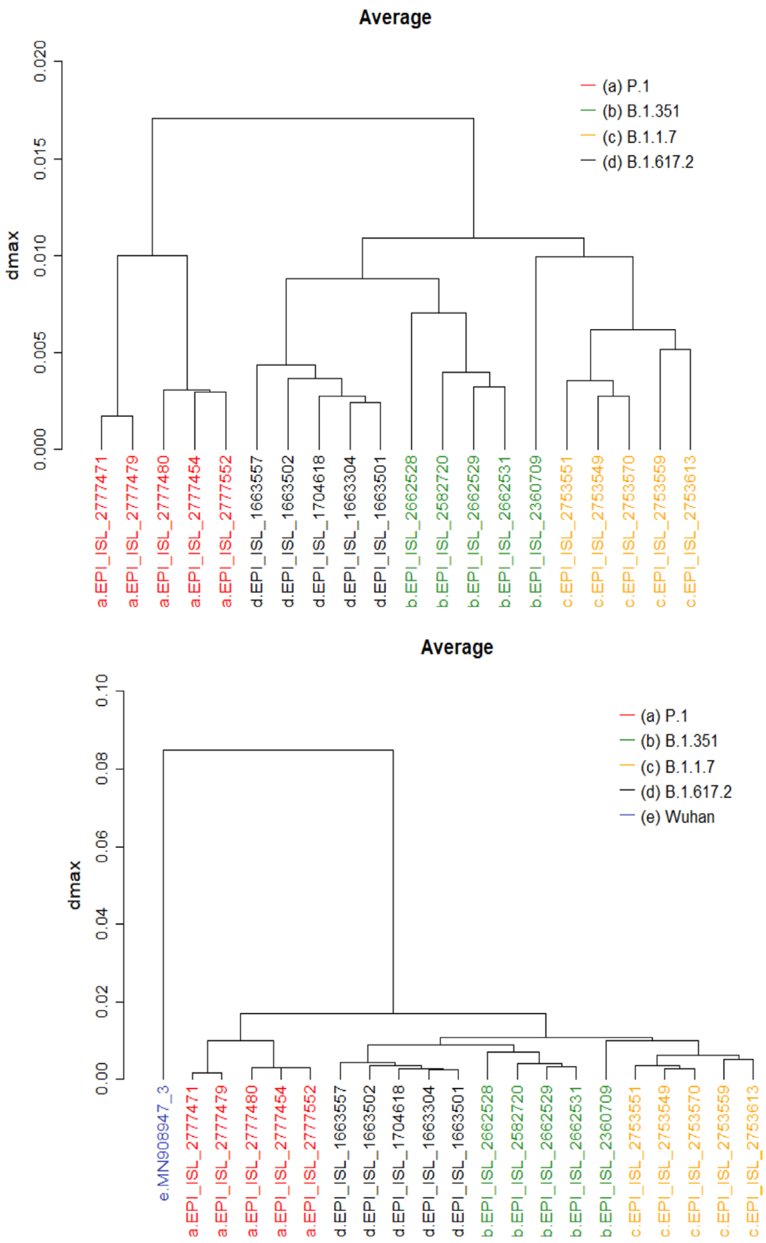

Fig. 1 Average type dendrogram build from the *dmax* values, see Definition 1-ii., using complete sequences of SARS-CoV 2, B.1.1.7, B.1.351, B.1.617.2, and P.1 variants, listed on top of Table 6, $\Lambda(A) = A \times A^o$, with $o = 3$ and $G = 9$. The five best representative sequences per variant plus the original sequence of SARS-CoV 2, MN908947 version 3

among the best BIC values, the value $G = 9$. Then, to formalize the notions 1 and 2, we used one of the best four models.

The question that arises is if the variants can be considered different from the perspective of Definition 1-ii. It is a desired characteristic, as it would allow connecting the medical characterization with an automatic tool based on stochastic processes. In Fig. 1 dendrograms (Kaufman and Rousseeuw 1990) of the *dmax* values are presented using the five most representative sequences per variant, above, and to compare with the original sequence (MN908947 version 3—China (Wuhan)), below is a dendrogram that includes it. In the case is used $\Lambda(A) = A \times A^o$, $o = 3$ and $G = 9$ (see Eq. (2)).

Figure 1 shows a clear separation between the group of variants and the original sequence MN908947 (bottom dendrogram). In addition, we see that the variants also form separated groups, with one possible exception between the variants B.1.1.7 (UK), and B.1.351 (South Africa). Note that the sequence EPI_ISL_2360709 that most closely approximates the group of sequences B.1.1.7 is the most representative of the variant B.1.351, according to the results of Table 6. As far as we know, U (Definition 2) is the only statistical tool that allows us to determine an order between samples of stochastic processes. Even more, in the present problem *dmax* has allowed to identify the groups by variants. Again, as far as we know, that is the only statistical tool associated with a metric (Definition 1), which allows the quantification of the proximity/discrepancy between samples of stochastic processes.

We have chosen $G = 9$ to consolidate the comparison study that we set out to carry out at the marginal level. We also use a formal criterion that has allowed us to select the five most representative sequences per variant. The next stage is to look for the divergences and similarities between the variants.

In Sect. 3.3, we present the results of a joint model, given by Definition 3-2. In other words, we are going to obtain a representation of the four variants. The variants may share transition probabilities in certain states and not share transition probabilities in other states of the state space. It is precisely this discrimination that we wish to determine.

3.3 Joint model between the variants

In this section, we use the notion given by Definition 3-2. (with $m = 4$, related to the four variants) to identify the commonalities and discrepancies between the variants B.1.1.7 (associated to the index $j = 1$), B.1.351 (with $j = 2$), B.1.617.2 (with $j = 3$), and P.1 (with $j = 4$). We look for these answers in the state space

$$\{1, 2, 3, 4\} \times \Lambda(A),$$

with $\Lambda(A) = A \times A^o$ defined by the Eq. (2), with, $o = 3$ and $G = 9$, and $A = \{a, c, g, t\}$.

In this situation, the states in $\{1, 2, 3, 4\} \times \Lambda(A)$ are denoted by $(j, (z, s))$. Thus, $(j, (z, s))$ indicates that state (z, s) is associated with the variant j, with $j = 1, 2, 3, 4$, and $(z, s) \in \Lambda(A)$.

Through the metric introduced in Cordeiro et al. (2020), it is possible to identify the partition $\hat{\mathbb{P}}$ on $\{1, 2, 3, 4\} \times \Lambda(A)$. 82 parts were identified in this case, the composition of the parts are displayed in Tables 12, 13, 14 and 15. The elements in each part indicate precisely which states of type $(j,(z,s)) \in \{1,...,4\} \times \Lambda(A)$ can be considered equivalent in the process of defining the transition probabilities, for each of the elements of space A. Without the assumptions of the Definition 3-2. there is a need to estimate one set of transition probabilities for each variant, and under the partition of $\{1,2,3,4\} \times \Lambda(A)$ we only need one set of transition probabilities. The BIC criterion points out which states should or should not be considered equivalent for the determination of the parts of the partition, and as a consequence which states will be used to estimate the same probability. Note that, if the variants are considered separately and without assuming any structure on the state space, there would be $4 \times 4^3 \times 3 = 768$ probabilities per variant, and, $4 \times 768 = 3072$ probabilities in total. Under the principle given by Definition 3-2. the number of probabilities drops to $82 \times 3 = 246$, this is around 8% of the original number of probabilities.

Tables 7 and 8 show the transition probabilities from each part (82 in total) of the state space to any element of the alphabet $A = \{a, c, g, t\}$, we identify the highest value per part in bold letter. Note that only five parts record transition probabilities superior to 0.5. Those are, 63, 68, 69, 76 and 78. As anticipated, it is first necessary to estimate $\mathbb{P}$ so that based on the identified parts, it is possible to calculate the probabilities given by the Eq. (8), which are those reported in Tables 7 and 8.

We place here the issue of indistinguishable versus distinguishable. That is, we seek in relation to which states of $\Lambda(A)$ the variants behave in an indistinguishable or distinguishable way. Two variants i and j will be distinguishable in relation to a specific state $(z,s) \in \Lambda(A)$ if such state is introduced in different parts in the set of 82 parts, then $(i,(z,s)) \in \hat{\mathcal{P}}_{i_{(z,s)}} \neq \hat{\mathcal{P}}_{j_{(z,s)}}$ and $(j,(z,s)) \in \hat{\mathcal{P}}_{j_{(z,s)}}$. This means that in relation to that state (z, s) the variants do not share the transition probabilities and therefore the states end up in different parts. That fact is established in Definition 3.

We build Table 9, which reports parts composed of (1) one state in $\Lambda(A)$, (2) two states in $\Lambda(A)$, (3) three states in $\Lambda(A)$, (4) four states in $\Lambda(A)$, (5) five states in $\Lambda(A)$, and (6) six states in $\Lambda(A)$. Table 9 reports elements of $\Lambda(A)$ for which all the variants are indistinguishable. In that table, for each element $(i, (z, s))$, we avoid the index i related to the variant since all four cases are inside the part. Then, the variants are indistinguishable concerning the states reported in that table. In other words, for such states, the processes that the variants represent are indistinguishable.

We see that of a total of 82 parts the variants fully agree in 62, they are detailed in Table 9. That is, we obtained nine parts with one element in $\Lambda(A)$, 18 parts with two elements in $\Lambda(A)$, 18 with three elements in $\Lambda(A)$, twelve with four elements in $\Lambda(A)$, two parts with five elements in $\Lambda(A)$, and three parts with six elements in $\Lambda(A)$. That is to say, of the 82 parts reported in the Tables 12, 13, 14 and 15, 20 parts reveal the discrepancies between the variants.

To rescue the discrepancies between the variants, we built Table 10, listing the 20 parts in which at least one discrepancy is found between the four variants. In Table 10, we have highlighted such cases in bold letter. For example, in part 38, we found the states (a, taa), (c, atc) for which the four variants coincide, but as for state (g, ctc), only variant 2 shares the transition probabilities with the other two states, and the remaining variants allocate the state (g, ctc) in part 60.

There are eleven states responsible for the discrepancies between the variants, marked in bold in Table 10. The discrepancies between the variants are related to

Table 7 Transition probabilities (see Eq. (8)), from each part $\hat{\mathcal{P}}_i, 1 \leq i \leq 42$, to each element of the alphabet $A = \{a, c, g, t\}$. In bold type, the highest probability per part

	$\hat{P}(a\|\hat{\mathcal{P}}_i)$	$\hat{P}(c\|\hat{\mathcal{P}}_i)$	$\hat{P}(g\|\hat{\mathcal{P}}_i)$	$\hat{P}(t\|\hat{\mathcal{P}}_i)$
$i = 1$	**0.301**	0.199	0.233	0.268
$i = 2$	**0.413**	0.192	0.073	0.322
$i = 3$	**0.399**	0.147	0.176	0.278
$i = 4$	0.173	0.150	0.284	**0.394**
$i = 5$	**0.393**	0.183	0.205	0.219
$i = 6$	**0.449**	0.106	0.074	0.371
$i = 7$	0.253	0.193	0.144	**0.410**
$i = 8$	0.274	0.120	0.200	**0.405**
$i = 9$	**0.341**	0.216	0.179	0.264
$i = 10$	0.270	0.185	0.086	**0.459**
$i = 11$	0.223	0.142	0.265	**0.369**
$i = 12$	**0.348**	0.196	0.229	0.227
$i = 13$	0.360	0.151	0.054	**0.436**
$i = 14$	**0.283**	0.167	0.272	0.279
$i = 15$	**0.326**	0.275	0.173	0.226
$i = 16$	**0.419**	0.221	0.076	0.283
$i = 17$	**0.374**	0.152	0.134	0.340
$i = 18$	0.232	0.206	0.213	**0.350**
$i = 19$	**0.326**	0.169	0.225	0.279
$i = 20$	**0.347**	0.308	0.170	0.174
$i = 21$	**0.338**	0.102	0.224	0.336
$i = 22$	**0.451**	0.090	0.174	0.285
$i = 23$	0.190	0.360	0.000	**0.450**
$i = 24$	0.270	0.167	**0.324**	0.239
$i = 25$	**0.334**	0.182	0.166	0.318
$i = 26$	**0.490**	0.133	0.099	0.278
$i = 27$	0.250	0.197	0.247	**0.306**
$i = 28$	**0.427**	0.158	0.055	0.360
$i = 29$	0.277	0.200	0.218	**0.305**
$i = 30$	0.213	0.156	**0.380**	0.251
$i = 31$	0.376	0.097	0.113	**0.414**
$i = 32$	0.338	0.196	0.086	**0.380**
$i = 33$	**0.353**	0.237	0.210	0.199
$i = 34$	0.301	0.230	0.114	**0.355**
$i = 35$	0.231	0.058	**0.465**	0.246
$i = 36$	0.257	0.234	0.197	**0.312**
$i = 37$	0.289	0.138	0.284	**0.290**
$i = 38$	**0.392**	0.195	0.112	0.301
$i = 39$	0.243	0.147	**0.320**	0.290
$i = 40$	0.125	0.259	0.243	**0.373**
$i = 41$	0.274	0.139	0.161	**0.426**
$i = 42$	0.231	0.160	0.146	**0.463**

Table 8 Transition probabilities (see Eq. (8)), from each part $\hat{\mathcal{P}}_i, 43 \leq i \leq 82$, to each element of the alphabet $A = \{a, c, g, t\}$. In bold type, the highest probability per part

	$\hat{P}(a\|\hat{\mathcal{P}}_i)$	$\hat{P}(c\|\hat{\mathcal{P}}_i)$	$\hat{P}(g\|\hat{\mathcal{P}}_i)$	$\hat{P}(t\|\hat{\mathcal{P}}_i)$
$i = 43$	0.235	0.155	0.209	**0.401**
$i = 44$	**0.317**	0.281	0.124	0.279
$i = 45$	0.212	**0.286**	0.218	0.283
$i = 46$	**0.467**	0.214	0.017	0.302
$i = 47$	0.196	0.139	0.319	**0.345**
$i = 48$	0.187	0.199	0.225	**0.390**
$i = 49$	**0.299**	0.242	0.197	0.262
$i = 50$	0.273	**0.301**	0.212	0.214
$i = 51$	0.283	0.229	0.038	**0.449**
$i = 52$	0.268	0.173	**0.391**	0.168
$i = 53$	**0.488**	0.042	0.115	0.355
$i = 54$	0.380	0.056	0.156	**0.409**
$i = 55$	0.165	0.126	**0.398**	0.310
$i = 56$	0.210	**0.311**	0.283	0.197
$i = 57$	0.255	0.110	0.293	**0.343**
$i = 58$	**0.338**	0.237	0.275	0.150
$i = 59$	0.261	0.264	0.136	**0.339**
$i = 60$	**0.385**	0.201	0.152	0.263
$i = 61$	0.374	0.150	0.081	**0.395**
$i = 62$	**0.281**	0.222	0.257	0.239
$i = 63$	0.265	0.135	0.067	**0.533**
$i = 64$	0.307	0.180	0.050	**0.463**
$i = 65$	0.210	0.241	0.170	**0.379**
$i = 66$	0.344	0.000	0.189	**0.467**
$i = 67$	0.151	0.194	0.165	**0.491**
$i = 68$	**0.658**	0.000	0.164	0.178
$i = 69$	0.078	0.108	**0.549**	0.265
$i = 70$	0.223	0.118	**0.362**	0.297
$i = 71$	0.223	**0.334**	0.120	0.324
$i = 72$	0.299	0.040	**0.342**	0.320
$i = 73$	0.191	0.298	**0.378**	0.133
$i = 74$	**0.470**	0.191	0.117	0.222
$i = 75$	**0.337**	0.154	0.260	0.249
$i = 76$	0.285	0.170	0.031	**0.514**
$i = 77$	0.297	0.150	0.228	**0.325**
$i = 78$	0.210	0.094	0.131	**0.565**
$i = 79$	0.292	0.067	0.260	**0.380**
$i = 80$	0.181	0.100	0.343	**0.376**
$i = 81$	**0.467**	0.121	0.024	0.389
$i = 82$	0.219	0.175	0.185	**0.422**

Table 9 Parts in which the variants are indistinguishable. From left to right, from top to bottom, composed of (1) one state, (2) two states, (3) three states, (4) four states, (5) five states, (6) six states. $\Lambda(A) = A \times A^o$, with $o = 3$ and $G = 9$ as given by Eq. (2). On the right of each part are the elements $(z, s) : (z, s) \in \Lambda(A)$ without the indicator of the variant, since all the four cases (one by variant) are in the part. See Tables 12, 13, 14 and 15

Part	State
$\hat{\mathcal{P}}_{35}$	(a, ggt)
$\hat{\mathcal{P}}_{66}$	(g, ccc)
$\hat{\mathcal{P}}_{68}$	(g, cgc)
$\hat{\mathcal{P}}_{69}$	(g, cgt)
$\hat{\mathcal{P}}_{73}$	(g, gta)
$\hat{\mathcal{P}}_{74}$	(g, tac)
$\hat{\mathcal{P}}_{79}$	(t, ggt)
$\hat{\mathcal{P}}_{80}$	(t, tat)
$\hat{\mathcal{P}}_{81}$	(t, tcc)

Part	States
$\hat{\mathcal{P}}_{1}$	(a, aaa), (a, atg)
$\hat{\mathcal{P}}_{4}$	(a, aat), (c, aat)
$\hat{\mathcal{P}}_{23}$	(a, cgg), (c, ccg)
$\hat{\mathcal{P}}_{28}$	(a, gac), (t, atc)
$\hat{\mathcal{P}}_{30}$	(a, gat), (g, agt)
$\hat{\mathcal{P}}_{41}$	(a, tct), (t, ctt)
$\hat{\mathcal{P}}_{48}$	(c, atg), (t, ttg)
$\hat{\mathcal{P}}_{51}$	(c, cga), (t, tgg)
$\hat{\mathcal{P}}_{52}$	(c, cgt), (g, gaa)
$\hat{\mathcal{P}}_{53}$	(c, ctc), (g, gcc)
$\hat{\mathcal{P}}_{59}$	(c, gtg), (t, cta)
$\hat{\mathcal{P}}_{64}$	(c, tgc), (t, tgc)
$\hat{\mathcal{P}}_{67}$	(g, ccg), (t, acg)
$\hat{\mathcal{P}}_{72}$	(g, ggt), (t, cgt)
$\hat{\mathcal{P}}_{75}$	(t, aag), (t, gaa)
$\hat{\mathcal{P}}_{77}$	(t, att), (t, ttt)
$\hat{\mathcal{P}}_{78}$	(t, ccc), (t, gcg)
$\hat{\mathcal{P}}_{82}$	(t, tct), (t, tgt)

Part	States
$\hat{\mathcal{P}}_{5}$	(a, aca), (a, gca), (t, tca)
$\hat{\mathcal{P}}_{8}$	(a, act), (c, tct), (t, gtt)
$\hat{\mathcal{P}}_{11}$	(a, agt), (t, aat), (t, agt)
$\hat{\mathcal{P}}_{18}$	(a, cat), (a, ctt), (t, gtg)
$\hat{\mathcal{P}}_{20}$	(a, ccg), (a, cgc), (g, ata)

Table 9 (continued)

Part	States
$\hat{\mathcal{P}}_{21}$	$(a, cct), (a, gct), (g, cct)$
$\hat{\mathcal{P}}_{22}$	$(a, cga), (c, ggc), (g, tcc)$
$\hat{\mathcal{P}}_{26}$	$(a, ctc), (c, gac), (t, gac)$
$\hat{\mathcal{P}}_{37}$	$(a, gtt), (g, tct), (g, ttt)$
$\hat{\mathcal{P}}_{39}$	$(a, tat), (g, act), (g, tgt)$
$\hat{\mathcal{P}}_{43}$	$(a, tgt), (c, tgt), (t, act)$
$\hat{\mathcal{P}}_{44}$	$(a, tta), (g, cta), (t, caa)$
$\hat{\mathcal{P}}_{55}$	$(c, gat), (g, aat), (t, gat)$
$\hat{\mathcal{P}}_{56}$	$(c, gca), (c, gta), (g, acg)$
$\hat{\mathcal{P}}_{58}$	$(c, gga), (g, cag), (g, gga)$
$\hat{\mathcal{P}}_{63}$	$(c, tcg), (g, cga), (g, tgc)$
$\hat{\mathcal{P}}_{70}$	$(g, gat), (g, gct), (t, gct)$
$\hat{\mathcal{P}}_{76}$	$(t, agc), (t, ggc), (t, gtc)$
Part	States
$\hat{\mathcal{P}}_{2}$	$(a, aac), (a, tac), (g, aac), (g, gtc)$
$\hat{\mathcal{P}}_{3}$	$(a, aag), (g, aag), (t, cag), (t, cga)$
$\hat{\mathcal{P}}_{12}$	$(a, ata), (a, tca), (g, aga), (t, aga)$
$\hat{\mathcal{P}}_{14}$	$(a, att), (a, ttt), (c, ctt), (g, aaa)$
$\hat{\mathcal{P}}_{15}$	$(a, caa), (c, ata), (c, gaa), (t, gca)$
$\hat{\mathcal{P}}_{24}$	$(a, cgt), (a, gaa), (g, ctt), (g, gtt)$
$\hat{\mathcal{P}}_{33}$	$(a, gga), (a, gta), (t, cca), (t, gga)$
$\hat{\mathcal{P}}_{40}$	$(a, tcg), (g, gag), (g, gcg), (t, cgg)$
$\hat{\mathcal{P}}_{47}$	$(c, agt), (c, gtt), (c, tat), (g, cat)$
$\hat{\mathcal{P}}_{50}$	$(c, cct), (g, caa), (g, gca), (t, gta)$
$\hat{\mathcal{P}}_{57}$	$(c, gct), (c, ggt), (g, gtg), (g, tat)$
$\hat{\mathcal{P}}_{62}$	$(c, tca), (g, aca), (g, cca), (t, atg)$
Part	States
$\hat{\mathcal{P}}_{6}$	$(a, acc), (a, ggc), (g, atc), (t, cgc), (t, ctc)$
$\hat{\mathcal{P}}_{10}$	$(a, agc), (a, tgc), (c, agc), (g, ggc), (t, ggg)$
Part	States
$\hat{\mathcal{P}}_{19}$	$(a, cca), (c, aga), (c, agg), (c, cag), (c, ttt), (t, gag)$
$\hat{\mathcal{P}}_{25}$	$(a, cta), (a, tga), (g, agg), (g, taa), (t, ata), (t, taa)$
$\hat{\mathcal{P}}_{34}$	$(a, ggg), (c, cac), (c, tgg), (g, cac), (g, cgg), (t, tga)$

Table 10 In bold letter elements for which the variants are different. $\Lambda(A) = A \times A^o$, with $o = 3$ and $G = 9$ as given by Eq. (2). On the right of each part are (a) (in bold) the elements (i, (z, s)) with $i = 1, \ldots, 4$ and $(z, s) \in \Lambda(A)$, $i = 1$(B.1.1.7 variant), $i = 2$(B.1.351 variant), $i = 3$(B.1.617.2 variant) and $i = 4$(P.1 variant) , (b) the elements $(z, s) : (z, s) \in \Lambda(A)$ without the indicator of the variant, since all the four cases (one by variant) are in the part. See Tables 12, 13, 14 and 15

$\hat{\mathcal{P}}_7$	$(a, acg), (a, agg), (c, acg), (g, tga), (g, tgg), (t, agg), (\mathbf{2}, (\mathbf{c}, \mathbf{ggg}))$
$\hat{\mathcal{P}}_9$	$(a, aga), (c, aag), (c, aca), (c, cta), (c, tga), (c, tta), (t, aca), (\mathbf{3}, (\mathbf{g}, \mathbf{tag}))$
$\hat{\mathcal{P}}_{13}$	$(a, atc), (\mathbf{1}, (\mathbf{a}, \mathbf{tcc})), (a, ttc), (t, tac), (t, ttc)$
$\hat{\mathcal{P}}_{16}$	$(a, cac), (a, ccc), (c, aac), (g, gac), (\mathbf{4}, (\mathbf{c}, \mathbf{acc}))$
$\hat{\mathcal{P}}_{17}$	$(a, cag), (c, gtc), (g, ttc), (t, tag), (\mathbf{3}, (\mathbf{a}, \mathbf{gtc})), (\mathbf{4}, (\mathbf{a}, \mathbf{gtc}))$
$\hat{\mathcal{P}}_{27}$	$(a, ctg), (c, act), (g, atg), (g, att), (t, cat), (\mathbf{2}, (\mathbf{c}, \mathbf{att}))$
$\hat{\mathcal{P}}_{29}$	$(a, gag), (a, ttg), (\mathbf{1}, (\mathbf{c}, \mathbf{att})), (t, aaa), (\mathbf{3}, (\mathbf{c}, \mathbf{att})), (\mathbf{4}, (\mathbf{c}, \mathbf{att}))$
$\hat{\mathcal{P}}_{31}$	$(a, gcc), (\mathbf{1}, (\mathbf{c}, \mathbf{ccc})), (c, cgc), (g, acc), (t, gcc), (\mathbf{2}, (\mathbf{c}, \mathbf{ccc})), (\mathbf{3}, (\mathbf{c}, \mathbf{ccc}))$
$\hat{\mathcal{P}}_{32}$	$(a, gcg), (\mathbf{1}, (\mathbf{a}, \mathbf{gtc})), (a, tag), (c, gcc), (c, tcc), (t, cac), (\mathbf{2}, (\mathbf{a}, \mathbf{gtc})), (\mathbf{2}, (\mathbf{t}, \mathbf{acc})),$ $(\mathbf{3}, (\mathbf{t}, \mathbf{acc})), (\mathbf{4}, (\mathbf{t}, \mathbf{acc}))$
$\hat{\mathcal{P}}_{36}$	$(a, gtg), (\mathbf{1}, (\mathbf{c}, \mathbf{cgg})), (c, ctg), (g, ttg), (t, tcg)$
$\hat{\mathcal{P}}_{38}$	$(a, taa), (c, atc), (\mathbf{2}, (\mathbf{g}, \mathbf{ctc}))$
$\hat{\mathcal{P}}_{42}$	$(a, tgg), (\mathbf{1}, (\mathbf{c}, \mathbf{ggg})), (t, cct), (\mathbf{3}, (\mathbf{c}, \mathbf{ggg})), (\mathbf{4}, (\mathbf{c}, \mathbf{ggg})),$
$\hat{\mathcal{P}}_{45}$	$(c, aaa), (c, cat), (c, cca), (c, gcg), (\mathbf{3}, (\mathbf{t}, \mathbf{ccg}))$
$\hat{\mathcal{P}}_{46}$	$(\mathbf{1}, (\mathbf{c}, \mathbf{acc})), (\mathbf{2}, (\mathbf{c}, \mathbf{acc})), (\mathbf{3}, (\mathbf{c}, \mathbf{acc}))$
$\hat{\mathcal{P}}_{49}$	$(c, caa), (\mathbf{1}, (\mathbf{g}, \mathbf{tag})), (g, tta), (\mathbf{2}, (\mathbf{g}, \mathbf{tag})), (\mathbf{4}, (\mathbf{g}, \mathbf{tag}))$
$\hat{\mathcal{P}}_{54}$	$(c, gag), (g, agc), (\mathbf{4}, (\mathbf{c}, \mathbf{ccc}))$
$\hat{\mathcal{P}}_{60}$	$(c, taa), (c, tag), (\mathbf{1}, (\mathbf{g}, \mathbf{ctc})), (g, tca), (\mathbf{3}, (\mathbf{g}, \mathbf{ctc})), (\mathbf{4}, (\mathbf{g}, \mathbf{ctc}))$
$\hat{\mathcal{P}}_{61}$	$(c, tac), (c, ttc), (t, aac), (\mathbf{1}, (\mathbf{t}, \mathbf{acc})), (\mathbf{2}, (\mathbf{a}, \mathbf{tcc})), (\mathbf{3}, (\mathbf{a}, \mathbf{tcc})), (\mathbf{4}, (\mathbf{a}, \mathbf{tcc}))$
$\hat{\mathcal{P}}_{65}$	$(c, ttg), (g, ctg), (t, ctg), (\mathbf{2}, (\mathbf{c}, \mathbf{cgg})), (\mathbf{2}, (\mathbf{t}, \mathbf{ccg})), (\mathbf{3}, (\mathbf{c}, \mathbf{cgg})), (\mathbf{4}, (\mathbf{c}, \mathbf{cgg})), (\mathbf{4}, (\mathbf{t}, \mathbf{ccg}))$
$\hat{\mathcal{P}}_{71}$	$(g, ggg), (g, tcg), (\mathbf{1}, (\mathbf{t}, \mathbf{ccg})), (t, tta)$

the transition probabilities shown in Tables 7, 8. We show such probabilities in Table 11, to visualize how the variants behave concerning such states.

The different magnitudes in the transition probabilities for each element of the alphabet A are those that indicate the divergence between the variants concerning each of these states. For example, if we focus on state (a, tcc), we see that the difference between variant 1 (B.1.1.7) and the other three variants (B.1.351, B.1.617.2, P.1) occurs to a greater or lesser extent in all transition probabilities (Table 11 on top) being that under variant 1 $\hat{P}(t|(a, tcc)) = 0.436$, already under any of the other three variants, such probability falls to $\hat{P}(t|(a, tcc)) = 0.395$.

Table 11 Transition probabilities from Tables 7 and 8 of the eleven elements of $\Lambda(A)$ where the variants diverge (parts in Table 10). In bold type, the highest probability per line

$(z, s) = (a, tcc)$

Variant	$P(a\|(z, s))$	$P(c\|(z, s))$	$P(g\|(z, s))$	$P(t\|(z, s))$
1 (B.1.1.7)	0.360	0.151	0.054	**0.436**
2, 3, 4 (B.1.351, B.1.617.2, P.1)	0.374	0.150	0.081	**0.395**

$(z, s) = (t, acc)$

Variant	$P(a\|(z, s))$	$P(c\|(z, s))$	$P(g\|(z, s))$	$P(t\|(z, s))$
1 (B.1.1.7)	0.374	0.150	0.081	**0.395**
2, 3, 4 (B.1.351, B.1.617.2, P.1)	0.338	0.196	0.086	**0.379**

$(z, s) = (c.cgg)$

Variant	$P(a\|(z, s))$	$P(c\|(z, s))$	$P(g\|(z, s))$	$P(t\|(z, s))$
1 (B.1.1.7)	0.257	0.234	0.197	**0.312**
2, 3, 4 (B.1.351, B.1.617.2, P.1)	0.210	0.241	0.170	**0.379**

$(z, s) = (t.ccg)$

Variant	$P(a\|(z, s))$	$P(c\|(z, s))$	$P(g\|(z, s))$	$P(t\|(z, s))$
1 (B.1.1.7)	0.223	**0.334**	0.120	0.324
2, 4 (B.1.351, P.1)	0.210	0.241	0.170	**0.380**
3 (B.1.617.2)	0.212	**0.286**	0.218	0.283

$(z, s) = (c, att)$

Variant	$P(a\|(z, s))$	$P(c\|(z, s))$	$P(g\|(z, s))$	$P(t\|(z, s))$
2 (B.1.351)	0.250	0.197	0.247	**0.306**
1, 3, 4 (B.1.1.7, B.1.617.2, P.1)	0.277	0.200	0.218	**0.305**

$(z, s) = (g, ctc)$

Variant	$P(a\|(z, s))$	$P(c\|(z, s))$	$P(g\|(z, s))$	$P(t\|(z, s))$
2 (B.1.351)	**0.392**	0.196	0.112	0.301
1, 3, 4 (B.1.1.7, B.1.617.2, P.1)	**0.385**	0.201	0.152	0.263

$(z, s) = (c, ggg)$

Variant	$P(a\|(z, s))$	$P(c\|(z, s))$	$P(g\|(z, s))$	$P(t\|(z, s))$
2 (B.1.351)	0.253	0.193	0.144	**0.410**
1, 3, 4 (B.1.1.7, B.1.617.2, P.1)	0.231	0.160	0.146	**0.463**

$(z, s) = (g, tag)$

Variant	$P(a\|(z, s))$	$P(c\|(z, s))$	$P(g\|(z, s))$	$P(t\|(z, s))$
3 (B.1.617.2)	**0.341**	0.216	0.179	0.264
1,2, 4 (B.1.1.7, B.1.351, P.1)e	**0.299**	0.242	0.197	0.262

Table 11 (continued)

$(z,s)=(c,acc)$

Variant	$P(a\|(z,s))$	$P(c\|(z,s))$	$P(g\|(z,s))$	$P(t\|(z,s))$
4 (P.1)	**0.419**	0.221	0.076	0.283
1,2, 3 (B.1.1.7, B.1.351, B.1.617.2)	**0.467**	0.214	0.017	0.302

$(z,s)=(c,ccc)$

Variant	$P(a\|(z,s))$	$P(c\|(z,s))$	$P(g\|(z,s))$	$P(t\|(z,s))$
4 (P.1)	0.380	0.056	0.156	**0.409**
1,2, 3 (B.1.1.7, B.1.351, B.1.617.2)	0.376	0.097	0.113	**0.414**

$(z,s)=(a,gtc)$

Variant	$P(a\|(z,s))$	$P(c\|(z,s))$	$P(g\|(z,s))$	$P(t\|(z,s))$
1,2 (B.1.1.7, B.1.351)	0.338	0.197	0.086	**0.379**
3,4 (B.1.617.2, P.1)	**0.374**	0.152	0.134	0.340

4 Discussion and conclusion

The data treated in this work are complete genomic sequences in *Fasta* format on the alphabet $A=\{a,c,g,t\}$, of SARS-CoV 2 virus. These are sequences coming from four variants, considered during the evolution of the pandemic as variants of concern. The variants investigated here are B.1.1.7 (UK), B.1.351 (South Africa), B.1.617.2 (India), and P.1 (Brazil). Table 1 and Appendix give the details of the database investigated in this paper as well as its source. The sequences (Table 1) are treated as samples coming from stochastic processes. Tools coming from stochastic processes are used on such sequences, see García and González-López (2017), García et al. (2018), Fernández et al. (2020), and Cordeiro et al. (2020). Such an assumption is supported by the fact that proteins are assembled by the concatenation of elements of the genetic alphabet A.

In Sect. 3.2, and through the Bayesian Information Criterion (BIC), a state-space $\Lambda(A)$ compatible with all sequences is identified, resulting in $\Lambda(A)=A\times A^o$, with $o=3$ and $G=9$. In that space act the transition probabilities given by Eq. (2). We first investigate the behavior of each sequence separately and using the proposal introduced in García et al. (2020a), which is the model given by Definition 3, with $m=1$, since such a model introduces a parsimonious strategy by defining the parameters that must be estimated (in the case, the ideal partition and the transition probabilities). Then, from those models, we extract the BIC values.

It should be noted that in this process of determination of the state-space, is identified the heterogeneous performance of the sequences, regarding the choice of G (see Tables 2, 3, 4, and 5) and $G=9$ is one of the four best options for all of them. By a stochastic classifier, we deduce that there are sequences, that best represent each variant. Such classification is identified from notion (2), see also Table 6 and Fernández et al. (2020) for details.

Using the best five sequences by variant, through the notion *dmax*, related to the metric given by Definition 1-i, see García et al. (2018), we build dendrograms, see Fig. 1, verifying the discrimination between the variants. Also, this analysis allows the discrimination between the sequences coming from the variants when compared with the original sequence of SARS-CoV 2 virus, MN908947 version 3.

With the selected sequences (five per variant, considered independent), we proceed to build a joint model between the variants. This model, given by Definition 3, represents the sets of samples through (a) a partition of the space $\{1, 2, 3, 4\} \times \Lambda(A)$ and (b) the transition probabilities of each part of the partition for each element of the alphabet A (Cordeiro et al. 2020). Denoting by $\{1, 2, 3, 4\}$ the collection of variants, $i = 1$(B.1.1.7 variant), $i = 2$(B.1.351 variant), $i = 3$(B.1.617.2 variant) and $i = 4$(P.1 variant) , the partition $\hat{\mathbb{P}}$ on $\{1, 2, 3, 4\} \times \Lambda(A)$ allows identifying the states (elements of $\Lambda(A)$) for which the variants show an indistinguishable performance, sharing their transition probabilities. Those cases are in Table 9, with probabilities reported in Tables 7 and 8. Thus, the commonalities between the variants are defined by the relation given by Definition 3-1. Such classification is carried out with the following idea behind, in the assembly of proteins there exist behavioral patterns shared by the four variants. We also show the 20 parts where the variants discriminate, see Table 10. It is necessary to note that a part reported in Table 10 can be composed of states for which the variants are indistinguishable and others showing different transition probabilities.

Eleven states out of a total of $4 \times 4^3 = 256$ states are responsible for the distinguishability between variants, see Table 11. Variant B.1.1.7 (UK) differs from the other three in the transition probabilities attributed to states (a, tcc), (t, acc), and (c, cgg). Variant B.1.351 (South Africa) differs from the other three in the transition probabilities attributed to states (c, att), (g, ctc), and (c, ggg). Variant B.1.617.2 (India) differs from the other three in the transition probabilities attributed to the state (g, tag). And variant P.1 (Brazil) differs from the other three in the transition probabilities attributed to states (c, acc), and (c, ccc). This evidence can characterize each of the variants relative to the other three variants. Two other states reveal discrepancies in subgroups of variants. State (t, ccg) separates variant B.1.1.7 (UK) from variant B.1.617.2 (India), and on the other hand, shows that variants B.1.351 (South Africa) and P.1 (Brazil) share the same transition probabilities. Already, for state (a, gtc), variants B.1.1.7 (UK) and B.1.351 (South Africa) are indistinguishable in terms of their transition probabilities, but, distinguishable from the group consisting of variants B.1.617.2 (India) and P.1 (Brazil).

In this paper, recently proposed methods and tools, within the framework of Markov processes, see García and González-López (2017), García et al. (2018), Fernández et al. (2020), and Cordeiro et al. (2020), are applied to the problem of characterization of four variants of the SARS-CoV 2 virus. The study that we decided to carry out is based on the assumption of identifying genetic sequences of the virus in *FASTA* format as being samples from stochastic processes. The novelties of this paper are: (a) to identify which genetic sequences best represent each variant from a set of sequences, (b) to produce a comparison between the variants recovering the discrimination between them, (c) to identify the states of the state space for which the variants are (c.1) indistinguishable, (c.2) distinguishable. In summary, we can affirm that through the stochastic representation offered by Definition 3, it is possible to represent the behavior of the variants and reveal the stochastic reasons for their discrepancies and similarities. We show how to use formal and recent tools in the area that go beyond models and allow us to measure and classify process' samples (in this case genetic sequences).

Appendix

Originating and submitting labs

Variant B.1.1.7, sequences: EPI_ISL_27535x, x = 44, 47, 49, 50, 51, 54, 55, 59, 63, 65, 69, 70, 72, 74, EPI_ISL_2753613, Originating lab: Respiratory Virus Unit, Microbiology Services Colindale, Public Health England (61 Colindale Avenue, London, NW9 5EQ, United Kingdom), Submitting lab: COVID-19 Genomics UK (COG-UK) Consortium (United Kingdom).

Variant B.1.351, sequences (1) EPI_ISL_2360687, EPI_ISL_2360709, EPI_ISL_2360747, Originating lab: Vaccines and Infectious Diseases Analytics Research Unit (VIDA) (1st Floor, Central West Wing, New Nurses Residence, Chris Hani Baragwanath Hospital, Soweto), Submitting lab: KRISP, KZn Research Innovation and Sequencing Platform (Nelson R Mandela School of Medicine, University of KwaZulu-natal, 719 Umbilo Road, Durban, South Africa). Sequence (2) EPI_ISL_2375982, Originating lab: Hanover Park CHC wc HPH (Cnr Surran Road &, Hanover Park Ave, Hanover Park, Cape Town, 7780, South Africa), Submitting lab: NHLS/UCT (Wellcome Centre for Infectious Disease Research in Africa, Institute of Infectious Disease and Molecular Medicine, Division of Medical Virology, Department of Pathology, Faculty of Health Sciences, University of Cape Town and National Health Laboratory Service, Anzio Rd Observatory 7925, Cape Town, South Africa). Sequence (3) EPI_ISL_2375983, Originating lab: Guguletu CHC wc GDH (Western Cape, Guguletu, Cape Town, 7750), Submitting lab: NHLS/UCT (Wellcome Centre for Infectious Disease Research in Africa, Institute of Infectious Disease and Molecular Medicine, Division of Medical Virology, Department of Pathology, Faculty of Health Sciences, University of Cape Town and National Health Laboratory Service, Anzio Rd Observatory 7925, Cape Town, South Africa). Sequences (4) EPI_ISL_2582710, EPI_ISL_2582720, Originating lab: PORT ELIZABETH LABORATORY (34 Ostrich Street, Southernwood, Mthatha 5099), Submitting lab: National Institute for Communicable Diseases of the National Health Laboratory Service (1 Modderfontein Road, Sandringham, Johannesburg, Gauteng, South Africa, 2131). Sequence (5) EPI_ISL_2621127, Originating lab: Groote Schuur Hospital wc GSH (Main Rd, Observatory, Cape Town, 7925, South Africa), Submitting lab: NHLS/UCT (Wellcome Centre for Infectious Disease Research in Africa, Institute of Infectious Disease and Molecular Medicine, Division of Medical Virology, Department of Pathology, Faculty of Health Sciences, University of Cape Town and National Health Laboratory Service, Anzio Rd Observatory 7925, Cape Town, South Africa). Sequences (6) EPI_ISL_2662514, EPI_ISL_2662518, EPI_ISL_2662528, Originating lab: DR GEORGE MUKHARI LABORATORY (Ga-Rankuwa, Gauteng, South Africa), Submitting lab: National Institute for Communicable Diseases of the National Health Laboratory Service (1 Modderfontein Road, Sandringham, Johannesburg, Gauteng, South Africa, 2131). Sequences (7) EPI_ISL_2753613, EPI_ISL_2753570, EPI_ISL_2753572, EPI_ISL_2753574, Originating lab: Respiratory Virus Unit, Microbiology Services Colindale, Public Health England (61 Colindale Avenue, London, NW9 5EQ, United Kingdom), Submitting lab: COVID-19 Genomics UK (COG-UK) Consortium (United Kingdom).

Variant B.1.617.2, sequence (1) EPI_ISL_1662291, Originating lab: Dept. Of Microbiology, Lt.. Baliram Kashyap Memorial Govt. Medical college, Dimrapal, Jagdalpur (Late Baliram Kashyap Memorial Government Medical College, Dimrapal, Jagdalpur, Chhattisgarh 494001), Submitting lab: Institute of Life Sciences - INSACOG (Institute of Life Sciences (ILS), NALCO Square, Bhubaneswar, Odisha, India, Pin Code: 751023). Sequences (2) EPI_ISL_1663304, EPI_ISL_1663312, Originating lab: CIIMS, Bilaspur, Chhattisgarh (Chhattisgarh - 495001), Submitting lab: Institute of Life Sciences - INSACOG (Institute of Life Sciences (ILS), NALCO Square, Bhubaneswar, Odisha, India, Pin Code: 751023). Sequence (3) EPI_ISL_1663376, Originating lab: NCCS, Pune (Ganeshkhind, Pune, Maharashtra 411007), Submitting lab: Institute of Life Sciences - INSACOG (Institute of Life Sciences (ILS), NALCO Square, Bhubaneswar, Odisha, India, Pin Code: 751023). Sequences (4) EPI_ISL_1663501, EPI_ISL_1663502, Originating lab: Veer Surendra Sai Institute of Medical Sciences and Research, Burla, Sambalpur (Odisha 768017), Submitting lab: Institute of Life Sciences - INSACOG (Institute of Life Sciences (ILS), NALCO Square, Bhubaneswar, Odisha, India, Pin Code: 751023). Sequences (5) EPI_ISL_1663507, EPI_ISL_1663522, Originating lab: AIIMS, Patna (Patna, Bihar 801507), Submitting lab: Institute of Life Sciences - INSACOG (Institute of Life Sciences (ILS), NALCO Square, Bhubaneswar, Odisha, India, Pin Code: 751023). Sequence (6) EPI_ISL_1663541, Originating lab: MGM Medical College, Jamshedpur (Jharkhand 831020), Submitting lab: Institute of Life Sciences - INSACOG (Institute of Life Sciences (ILS), NALCO Square, Bhubaneswar, Odisha, India, Pin Code: 751023). Sequences (7) EPI_ISL_16635x, x=57, 63 and 64, Originating lab: Immunogenomics lab, Institute of Life Sciences, Bhubaneswar (Institute of Life Sciences (ILS), NALCO Square, Bhubaneswar, Odisha, India, Pin Code: 751023), Submitting lab: Institute of Life Sciences - INSACOG (Institute of Life Sciences (ILS), NALCO Square, Bhubaneswar, Odisha, India, Pin Code: 751023). Sequences (8) EPI_ISL_1704x, x = 234, 618, 629, Originating lab: ICMR-National Institute of Virology - INSACOG (National Institute of Virology, 20-A Dr. Ambedkar Road, Post box-11, Pune-411001, Maharashtra, India), Submitting lab: NIV Influenza (National Institute of Virology, 20-A Dr. Ambedkar Road, Post box-11, Pune-411001, Maharashtra, India).

Variant P.1, sequences (1) EPI_ISL_27774x, x = 05, 06, 14, 16, 18, 54, 70, 71, 77, 78, 79, 80, Originating/Submitting lab: Laboratorio de Ecologia de Doencas Transmissiveis na Amazonia, Instituto Leonidas e Maria Deane - Fiocruz Amazonia (476 Terezina St, Adrianopolis, Manaus, Amazonas, Brazil. ZIP Code 69057-070). Sequences (2) EPI_ISL_2777507, EPI_ISL_2777509, and EPI_ISL_2777552, Originating lab: Laboratório Central de Saúde Pública do Amazonas - LACEN-AM (528 Emílio Moreira St, Manaus, Amazonas, Brazil. ZIP Code 69020-040), Submitting lab: Laboratorio de Ecologia de Doencas Transmissiveis na Amazonia, Instituto Leonidas e Maria Deane - Fiocruz Amazonia (476 Terezina St, Adrianopolis, Manaus, Amazonas, Brazil. ZIP Code 69057-070).

Table 12 Composition of parts from $\hat{\mathcal{P}}_1$ to $\hat{\mathcal{P}}_{21}$. On the right are the states (by part) $(i, (z, s))$ with $i = 1, ..., 4$ and $(z, s) \in \Lambda(A) = A \times A^o$, with $o = 3$ and $G = 9$ as given by Eq. (2). i = 1(B.1.1.7 variant), i = 2(B.1.351 variant), i = 3(B.1.617.2 variant) and i = 4(P.1 variant)

Part	States
$\hat{\mathcal{P}}_1$	$(1, (a, aaa)), (1, (a, atg)), (2, (a, aaa)), (2, (a, atg)), (3, (a, aaa)), (3, (a, atg)), (4, (a, aaa)), (4, (a, atg))$
$\hat{\mathcal{P}}_2$	$(1, (a, aac)), (1, (a, tac)), (1, (g, aac)), (1, (g, gtc)), (2, (a, aac)), (2, (a, tac)), (2, (g, aac)), (2, (g, gtc)), (3, (a, aac)), (3, (a, tac)), (3, (g, aac)), (3, (g, gtc)), (4, (a, aac)), (4, (a, tac)), (4, (g, aac)), (4, (g, gtc))$
$\hat{\mathcal{P}}_3$	$(1, (a, aag)), (1, (g, aag)), (1, (t, cag)), (1, (t, cga)), (2, (a, aag)), (2, (g, aag)), (2, (t, cag)), (2, (t, cga)), (3, (a, aag)), (3, (g, aag)), (3, (t, cag)), (3, (t, cga)), (4, (a, aag))(4, (g, aag)), (4, (t, cag)), (4, (t, cga))$
$\hat{\mathcal{P}}_4$	$(1, (a, aat)), (1, (c, aat)), (2, (a, aat)), (2, (c, aat)), (3, (a, aat)), (3, (c, aat)), (4, (a, aat)), (4, (c, aat))$
$\hat{\mathcal{P}}_5$	$(1, (a, aca)), (1, (a, gca)), (1, (t, tca)), (2, (a, aca)), (2, (a, gca)), (2, (t, tca)), (3, (a, aca)), (3, (a, gca)), (3, (t, tca)), (4, (a, aca)), (4, (a, gca)), (4, (t, tca))$
$\hat{\mathcal{P}}_6$	$(1, (a, acc)), (1, (a, ggc)), (1, (g, atc)), (1, (t, cgc)), (1, (t, ctc)), (2, (a, acc)), (2, (a, ggc)), (2, (g, atc)), (2, (t, cgc)), (2, (t, ctc)), (3, (a, acc)), (3, (a, ggc)), (3, (g, atc)), (3, (t, cgc)), (3, (t, ctc)), (4, (a, acc)), (4, (a, ggc)), (4, (g, atc)), (4, (t, cgc)), (4, (t, ctc))$
$\hat{\mathcal{P}}_7$	$(1, (a, acg)), (1, (a, agg)), (1, (c, acg)), (1, (g, tga)), (1, (g, tgg)), (1, (t, agg)), (2, (a, acg)), (2, (c, acg)), (2, (a, agg)), (2, (c, ggg)), (2, (g, tga)), (2, (g, tgg)), (2, (t, agg)), (3, (a, acg)), (3, (a, agg)), (3, (c, acg)), (3, (g, tga)), (3, (g, tgg)), (3, (t, agg)), (4, (a, acg)), (4, (a, agg)), (4, (c, acg)), (4, (g, tga)), (4, (g, tgg)), (4, (t, agg))$
$\hat{\mathcal{P}}_8$	$(1, (a, act)), (1, (c, tct)), (1, (t, gtt)), (2, (a, act)), (2, (c, tct)), (2, (t, gtt)), (3, (a, act)), (3, (c, tct)), (3, (t, gtt)), (4, (a, act)), (4, (c, tct)), (4, (t, gtt))$
$\hat{\mathcal{P}}_9$	$(1, (a, aga)), (1, (c, aag)), (1, (c, aca)), (1, (c, cta)), (1, (c, tga)), (1, (c, tta)), (1, (t, aca)), (2, (a, aga)), (2, (c, aag)), (2, (c, aca)), (2, (c, cta)), (2, (c, tga)), (2, (c, tta)), (2, (t, aca)), (3, (a, aga)), (3, (c, aag)), (3, (c, aca)), (3, (c, cta)), (3, (c, tga)), (3, (c, tta)), (3, (g, tag)), (3, (t, aca)), (4, (a, aga)), (4, (c, aag)), (4, (c, aca)), (4, (c, cta)), (4, (c, tga)), (4, (c, tta)), (4, (t, aca))$
$\hat{\mathcal{P}}_{10}$	$(1, (a, agc)), (1, (a, tgc)), (1, (c, agc)), (1, (g, ggc)), (1, (t, ggg)), (2, (a, agc)), (2, (a, tgc)), (2, (c, agc)), (2, (g, ggc)), (2, (t, ggg)), (3, (a, agc)), (3, (a, tgc)), (3, (c, agc)), (3, (g, ggc)), (3, (t, ggg)), (4, (a, agc)), (4, (a, tgc)), (4, (c, agc)), (4, (g, ggc)), (4, (t, ggg))$
$\hat{\mathcal{P}}_{11}$	$(1, (a, agt)), (1, (t, aat)), (1, (t, agt)), (2, (a, agt)), (2, (t, aat)), (2, (t, agt)), (3, (a, agt)), (3, (t, aat)), (3, (t, agt)), (4, (a, agt)), (4, (t, aat)), (4, (t, agt))$
$\hat{\mathcal{P}}_{12}$	$(1, (a, ata)), (1, (a, tca)), (1, (g, aga)), (1, (t, aga)), (2, (a, ata)), (2, (a, tca)), (2, (g, aga)), (2, (t, aga)), (3, (a, ata)), (3, (a, tca)), (3, (g, aga)), (3, (t, aga)), (4, (a, ata)), (4, (a, tca)), (4, (g, aga)), (4, (t, aga))$
$\hat{\mathcal{P}}_{13}$	$(1, (a, atc)), (1, (a, tcc)), (1, (a, ttc)), (1, (t, tac)), (1, (t, ttc)), (2, (a, atc)), (2, (a, ttc)), (2, (t, tac)), (2, (t, ttc)), (3, (a, atc)), (3, (a, ttc)), (3, (t, tac)), (3, (t, ttc)), (4, (a, atc)), (4, (a, ttc)), (4, (t, tac)), (4, (t, ttc))$
$\hat{\mathcal{P}}_{14}$	$(1, (a, att)), (1, (a, ttt)), (1, (c, ctt)), (1, (g, aaa)), (2, (a, att)), (2, (a, ttt)), (2, (c, ctt)), (2, (g, aaa)), (3, (a, att)), (3, (a, ttt)), (3, (c, ctt)), (3, (g, aaa)), (4, (a, att)), (4, (a, ttt)), (4, (c, ctt)), (4, (g, aaa))$
$\hat{\mathcal{P}}_{15}$	$(1, (a, caa)), (1, (c, ata)), (1, (c, gaa)), (1, (t, gca)), (2, (a, caa)), (2, (c, ata)), (2, (c, gaa)), (2, (t, gca)), (3, (a, caa)), (3, (c, ata)), (3, (c, gaa)), (3, (t, gca)), (4, (a, caa)), (4, (c, ata)), (4, (c, gaa)), (4, (t, gca))$
$\hat{\mathcal{P}}_{16}$	$(1, (a, cac)), (1, (a, ccc)), (1, (c, aac)), (1, (g, gac)), (2, (a, cac)), (2, (a, ccc)), (2, (c, aac)), (2, (g, gac)), (3, (a, cac)), (3, (a, ccc)), (3, (c, aac)), (3, (g, gac)), (4, (a, cac)), (4, (a, ccc)), (4, (c, aac)), (4, (c, acc)), (4, (g, gac))$
$\hat{\mathcal{P}}_{17}$	$(1, (a, cag)), (1, (c, gtc)), (1, (g, ttc)), (1, (t, tag)), (2, (a, cag)), (2, (c, gtc)), (2, (g, ttc)), (2, (t, tag)), (3, (a, cag)), (3, (a, gtc)), (3, (c, gtc)), (3, (g, ttc)), (3, (t, tag)), (4, (a, cag)), (4, (a, gtc)), (4, (c, gtc)), (4, (g, ttc)), (4, (t, tag))$

Table 12 (continued)

Part	States
$\hat{\mathcal{P}}_{18}$	(1, (*a*, *cat*)), (1, (*a*, *ctt*)), (1, (*t*, *gtg*)), (2, (*a*, *cat*)), (2, (*a*, *ctt*)), (2, (*t*, *gtg*)), (3, (*a*, *cat*)), (3, (*a*, *ctt*)), (3, (*t*, *gtg*)), (4, (*a*, *cat*)), (4, (*a*, *ctt*)), (4, (*t*, *gtg*))
$\hat{\mathcal{P}}_{19}$	(1, (*a*, *cca*)), (1, (*c*, *aga*)), (1, (*c*, *agg*)), (1, (*c*, *cag*)), (1, (*c*, *ttt*)), (1, (*t*, *gag*)), (2, (*a*, *cca*)), (2, (*c*, *aga*)), (2, (*c*, *agg*)), (2, (*c*, *cag*)), (2, (*c*, *ttt*)), (2, (*t*, *gag*)), (3, (*a*, *cca*)), (3, (*c*, *aga*)), (3, (*c*, *agg*)), (3, (*c*, *cag*)), (3, (*c*, *ttt*)), (3, (*t*, *gag*)), (4, (*a*, *cca*)), (4, (*c*, *aga*)), (4, (*c*, *agg*)), (4, (*c*, *cag*)), (4, (*c*, *ttt*)), (4, (*t*, *gag*))
$\hat{\mathcal{P}}_{20}$	(1, (*a*, *ccg*)), (1, (*a*, *cgc*)), (1, (*g*, *ata*)), (2, (*a*, *ccg*)), (2, (*a*, *cgc*)), (2, (*g*, *ata*)), (3, (*a*, *ccg*)), (3, (*a*, *cgc*)), (3, (*g*, *ata*)), (4, (*a*, *ccg*)), (4, (*a*, *cgc*)), (4, (*g*, *ata*))
$\hat{\mathcal{P}}_{21}$	(1, (*a*, *cct*)), (1, (*a*, *gct*)), (1, (*g*, *cct*)), (2, (*a*, *cct*)), (2, (*a*, *gct*)), (2, (*g*, *cct*)), (3, (*a*, *cct*)), (3, (*a*, *gct*)), (3, (*g*, *cct*)), (4, (*a*, *cct*)), (4, (*a*, *gct*)), (4, (*g*, *cct*))

Table 13 Composition of parts from $\hat{\mathcal{P}}_{22}$ to $\hat{\mathcal{P}}_{42}$. On the right are the states (by part) $(i, (z, s))$ with $i = 1, ..., 4$ and $(z, s) \in \Lambda(A) = A \times A^o$, with $o = 3$ and $G = 9$ as given by Eq. (2). i = 1(B.1.1.7 variant), i = 2(B.1.351 variant), i = 3(B.1.617.2 variant) and i = 4(P.1 variant)

Part	States
$\hat{\mathcal{P}}_{22}$	(1, (*a*, *cga*)), (1, (*c*, *ggc*)), (1, (*g*, *tcc*)), (2, (*a*, *cga*)), (2, (*c*, *ggc*)), (2, (*g*, *tcc*)), (3, (*a*, *cga*)), (3, (*c*, *ggc*)), (3, (*g*, *tcc*)), (4, (*a*, *cga*)), (4, (*c*, *ggc*)), (4, (*g*, *tcc*))
$\hat{\mathcal{P}}_{23}$	(1, (*a*, *cgg*)), (1, (*c*, *ccg*)), (2, (*a*, *cgg*)), (2, (*c*, *ccg*)), (3, (*a*, *cgg*)), (3, (*c*, *ccg*)), (4, (*a*, *cgg*)), (4, (*c*, *ccg*))
$\hat{\mathcal{P}}_{24}$	(1, (*a*, *cgt*)), (1, (*a*, *gaa*)), (1, (*g*, *ctt*)), (1, (*g*, *gtt*)), (2, (*a*, *cgt*)), (2, (*a*, *gaa*)), (2, (*g*, *ctt*)), (2, (*g*, *gtt*)), (3, (*a*, *cgt*)), (3, (*a*, *gaa*)), (3, (*g*, *ctt*)), (3, (*g*, *gtt*)), (4, (*a*, *cgt*)), (4, (*a*, *gaa*)), (4, (*g*, *ctt*)), (4, (*g*, *gtt*))
$\hat{\mathcal{P}}_{25}$	(1, (*a*, *cta*)), (1, (*a*, *tga*)), (1, (*g*, *agg*)), (1, (*g*, *taa*)), (1, (*t*, *ata*)), (1, (*t*, *taa*)), (2, (*a*, *cta*)), (2, (*a*, *tga*)), (2, (*g*, *agg*)), (2, (*g*, *taa*)), (2, (*t*, *ata*)), (2, (*t*, *taa*)), (3, (*a*, *cta*)), (3, (*a*, *tga*)), (3, (*g*, *agg*)), (3, (*g*, *taa*)), (3, (*t*, *ata*)), (3, (*t*, *taa*)), (4, (*a*, *cta*)), (4, (*a*, *tga*)), (4, (*g*, *agg*)), (4, (*g*, *taa*)), (4, (*t*, *ata*)), (4, (*t*, *taa*))
$\hat{\mathcal{P}}_{26}$	(1, (*a*, *ctc*)), (1, (*c*, *gac*)), (1, (*t*, *gac*)), (2, (*a*, *ctc*)), (2, (*c*, *gac*)), (2, (*t*, *gac*)), (3, (*a*, *ctc*)), (3, (*c*, *gac*)), (3, (*t*, *gac*)), (4, (*a*, *ctc*)), (4, (*c*, *gac*)), (4, (*t*, *gac*))
$\hat{\mathcal{P}}_{27}$	(1, (*a*, *ctg*)), (1, (*c*, *act*)), (1, (*g*, *atg*)), (1, (*g*, *att*)), (1, (*t*, *cat*)), (2, (*a*, *ctg*)), (2, (*c*, *act*)), (2, (*c*, *att*)), (2, (*g*, *atg*)), (2, (*g*, *att*)), (2, (*t*, *cat*)), (3, (*a*, *ctg*)), (3, (*c*, *act*)), (3, (*g*, *atg*)), (3, (*g*, *att*)), (3, (*t*, *cat*)), (4, (*a*, *ctg*)), (4, (*c*, *act*)), (4, (*g*, *atg*)), (4, (*g*, *att*)), (4, (*t*, *cat*))
$\hat{\mathcal{P}}_{28}$	(1, (*a*, *gac*)), (1, (*t*, *atc*)), (2, (*a*, *gac*)), (2, (*t*, *atc*)), (3, (*a*, *gac*)), (3, (*t*, *atc*)), (4, (*a*, *gac*)), (4, (*t*, *atc*))
$\hat{\mathcal{P}}_{29}$	(1, (*a*, *gag*)), (1, (*a*, *ttg*)), (1, (*c*, *att*)), (1, (*t*, *aaa*)), (2, (*a*, *gag*)), (2, (*a*, *ttg*)), (2, (*t*, *aaa*)), (3, (*a*, *gag*)), (3, (*a*, *ttg*)), (3, (*c*, *att*)), (3, (*t*, *aaa*)), (4, (*a*, *gag*)), (4, (*a*, *ttg*)), (4, (*c*, *att*)), (4, (*t*, *aaa*))
$\hat{\mathcal{P}}_{30}$	(1, (*a*, *gat*)), (1, (*g*, *agt*)), (2, (*a*, *gat*)), (2, (*g*, *agt*)), (3, (*a*, *gat*)), (3, (*g*, *agt*)), (4, (*a*, *gat*)), (4, (*g*, *agt*))
$\hat{\mathcal{P}}_{31}$	(1, (*a*, *gcc*)), (1, (*c*, *ccc*)), (1, (*c*, *cgc*)), (1, (*g*, *acc*)), (1, (*t*, *gcc*)), (2, (*a*, *gcc*)), (2, (*c*, *ccc*)), (2, (*c*, *cgc*)),

Table 13 (continued)

Part	States
	(2, (g, acc)), (2, (t, gcc)), (3, (a, gcc)), (3, (c, ccc)), (3, (c, cgc)), (3, (g, acc)), (3, (t, gcc)), (4, (a, gcc)),
	(4, (c, cgc)), (4, (g, acc)), (4, (t, gcc))
$\hat{\mathcal{P}}_{32}$	(1, (a, gcg)), (1, (a, gtc)), (1, (a, tag)), (1, (c, gcc)), (1, (c, tcc)), (1, (t, cac)), (2, (a, gcg)), (2, (a, gtc)),
	(2, (a, tag)), (2, (c, gcc)), (2, (c, tcc)), (2, (t, acc)), (2, (t, cac)), (3, (a, gcg)), (3, (a, tag)), (3, (c, gcc)),
	(3, (c, tcc)), (3, (t, acc)), (3, (t, cac)), (4, (a, gcg)), (4, (a, tag)), (4, (c, gcc)), (4, (c, tcc)), (4, (t, acc)),
	(4, (t, cac))
$\hat{\mathcal{P}}_{33}$	(1, (a, gga)), (1, (a, gta)), (1, (t, cca)), (1, (t, gga)), (2, (a, gga)), (2, (a, gta)), (2, (t, cca)), (2, (t, gga)),
	(3, (a, gga)), (3, (a, gta)), (3, (t, cca)), (3, (t, gga)), (4, (a, gga)), (4, (a, gta)), (4, (t, cca)), (4, (t, gga))
$\hat{\mathcal{P}}_{34}$	(1, (a, ggg)), (1, (c, cac)), (1, (c, tgg)), (1, (g, cac)), (1, (g, cgg)), (1, (t, tga)), (2, (a, ggg)), (2, (c, cac)),
	(2, (c, tgg)), (2, (g, cac)), (2, (g, cgg)), (2, (t, tga)), (3, (a, ggg)), (3, (c, cac)), (3, (c, tgg)), (3, (g, cac)),
	(3, (g, cgg)), (3, (t, tga)), (4, (a, ggg)), (4, (c, cac)), (4, (c, tgg)), (4, (g, cac)), (4, (g, cgg)), (4, (t, tga))
$\hat{\mathcal{P}}_{35}$	(1, (a, ggt)), (2, (a, ggt)), (3, (a, ggt)), (4, (a, ggt))
$\hat{\mathcal{P}}_{36}$	(1, (a, gtg)), (1, (c, cgg)), (1, (c, ctg)), (1, (g, ttg)), (1, (t, tcg)), (2, (a, gtg)), (2, (c, ctg)), (2, (g, ttg)),
	(2, (t, tcg)), (3, (a, gtg)), (3, (c, ctg)), (3, (g, ttg)), (3, (t, tcg)), (4, (a, gtg)), (4, (c, ctg)), (4, (g, ttg)),
	(4, (t, tcg))
$\hat{\mathcal{P}}_{37}$	(1, (a, gtt)), (1, (g, tct)), (1, (g, ttt)), (2, (a, gtt)), (2, (g, tct)), (2, (g, ttt)), (3, (a, gtt)), (3, (g, tct)),
	(3, (g, ttt)), (4, (a, gtt)), (4, (g, tct)), (4, (g, ttt))
$\hat{\mathcal{P}}_{38}$	(1, (a, taa)), (1, (c, atc)), (2, (a, taa)), (2, (c, atc)), (2, (g, ctc)), (3, (a, taa)), (3, (c, atc)), (4, (a, taa)),
	(4, (c, atc))
$\hat{\mathcal{P}}_{39}$	(1, (a, tat)), (1, (g, act)), (1, (g, tgt)), (2, (a, tat)), (2, (g, act)), (2, (g, tgt)), (3, (a, tat)), (3, (g, act)),
	(3, (g, tgt)), (4, (a, tat)), (4, (g, act)), (4, (g, tgt))
$\hat{\mathcal{P}}_{40}$	(1, (a, tcg)), (1, (g, gag)), (1, (g, gcg)), (1, (t, cgg)), (2, (a, tcg)), (2, (g, gag)), (2, (g, gcg)), (2, (t, cgg)),
	(3, (a, tcg)), (3, (g, gag)), (3, (g, gcg)), (3, (t, cgg)), (4, (a, tcg)), (4, (g, gag)), (4, (g, gcg)), (4, (t, cgg))
$\hat{\mathcal{P}}_{41}$	(1, (a, tct)), (1, (t, ctt)), (2, (a, tct)), (2, (t, ctt)), (3, (a, tct)), (3, (t, ctt)), (4, (a, tct)), (4, (t, ctt))
$\hat{\mathcal{P}}_{42}$	(1, (a, tgg)), (1, (c, ggg)), (1, (t, cct)), (2, (a, tgg)), (2, (t, cct)), (3, (a, tgg)), (3, (c, ggg)), (3, (t, cct)),
	(4, (a, tgg)), (4, (c, ggg)), (4, (t, cct))

Table 14 Composition of parts from $\hat{\mathcal{P}}_{43}$ to $\hat{\mathcal{P}}_{62}$. On the right are the states (by part) $(i, (z, s))$ with $i = 1, ..., 4$ and $(z, s) \in \Lambda(A) = A \times A^o$, with $o = 3$ and $G = 9$ as given by Eq. (2). $i = 1$(B.1.1.7 variant), $i = 2$(B.1.351 variant), $i = 3$(B.1.617.2 variant) and $i = 4$(P.1 variant)

Part	States
$\hat{\mathcal{P}}_{43}$	(1, (*a*, *tgt*)), (1, (*c*, *tgt*)), (1, (*t*, *act*)), (2, (*a*, *tgt*)), (2, (*c*, *tgt*)), (2, (*t*, *act*)), (3, (*a*, *tgt*)), (3, (*c*, *tgt*)), (3, (*t*, *act*)), (4, (*a*, *tgt*)), (4, (*c*, *tgt*)), (4, (*t*, *act*))
$\hat{\mathcal{P}}_{44}$	(1, (*a*, *tta*)), (1, (*g*, *cta*)), (1, (*t*, *caa*)), (2, (*a*, *tta*)), (2, (*g*, *cta*)), (2, (*t*, *caa*)), (3, (*a*, *tta*)), (3, (*g*, *cta*)), (3, (*t*, *caa*)), (4, (*a*, *tta*)), (4, (*g*, *cta*)), (4, (*t*, *caa*))
$\hat{\mathcal{P}}_{45}$	(1, (*c*, *aaa*)), (1, (*c*, *cat*)), (1, (*c*, *cca*)), (1, (*c*, *gcg*)), (2, (*c*, *aaa*)), (2, (*c*, *cat*)), (2, (*c*, *cca*)), (2, (*c*, *gcg*)), (3, (*c*, *aaa*)), (3, (*c*, *cat*)), (3, (*c*, *cca*)), (3, (*c*, *gcg*)), (3, (*t*, *ccg*)), (4, (*c*, *aaa*)), (4, (*c*, *cat*)), (4, (*c*, *cca*)), (4, (*c*, *gcg*))
$\hat{\mathcal{P}}_{46}$	(1, (*c*, *acc*)), (2, (*c*, *acc*)), (3, (*c*, *acc*))
$\hat{\mathcal{P}}_{47}$	(1, (*c*, *agt*)), (1, (*c*, *gtt*)), (1, (*c*, *tat*)), (1, (*g*, *cat*)), (2, (*c*, *agt*)), (2, (*c*, *gtt*)), (2, (*c*, *tat*)), (2, (*g*, *cat*)), (3, (*c*, *agt*)), (3, (*c*, *gtt*)), (3, (*c*, *tat*)), (3, (*g*, *cat*)), (4, (*c*, *agt*)), (4, (*c*, *gtt*)), (4, (*c*, *tat*)), (4, (*g*, *cat*))
$\hat{\mathcal{P}}_{48}$	(1, (*c*, *atg*)), (1, (*t*, *ttg*)), (2, (*c*, *atg*)), (2, (*t*, *ttg*)), (3, (*c*, *atg*)), (3, (*t*, *ttg*)), (4, (*c*, *atg*)), (4, (*t*, *ttg*))
$\hat{\mathcal{P}}_{49}$	(1, (*c*, *caa*)), (1, (*g*, *tag*)), (1, (*g*, *tta*)), (2, (*c*, *caa*)), (2, (*g*, *tag*)), (2, (*g*, *tta*)), (3, (*c*, *caa*)), (3, (*g*, *tta*)), (4, (*c*, *caa*)), (4, (*g*, *tag*)), (4, (*g*, *tta*))
$\hat{\mathcal{P}}_{50}$	(1, (*c*, *cct*)), (1, (*g*, *caa*)), (1, (*g*, *gca*)), (1, (*t*, *gta*)), (2, (*c*, *cct*)), (2, (*g*, *caa*)), (2, (*g*, *gca*)), (2, (*t*, *gta*)), (3, (*c*, *cct*)), (3, (*g*, *caa*)), (3, (*g*, *gca*)), (3, (*t*, *gta*)), (4, (*c*, *cct*)), (4, (*g*, *caa*)), (4, (*g*, *gca*)), (4, (*t*, *gta*))
$\hat{\mathcal{P}}_{51}$	(1, (*c*, *cga*)), (1, (*t*, *tgg*)), (2, (*c*, *cga*)), (2, (*t*, *tgg*)), (3, (*c*, *cga*)), (3, (*t*, *tgg*)), (4, (*c*, *cga*)), (4, (*t*, *tgg*))
$\hat{\mathcal{P}}_{52}$	(1, (*c*, *cgt*)), (1, (*g*, *gaa*)), (2, (*c*, *cgt*)), (2, (*g*, *gaa*)), (3, (*c*, *cgt*)), (3, (*g*, *gaa*)), (4, (*c*, *cgt*)), (4, (*g*, *gaa*))
$\hat{\mathcal{P}}_{53}$	(1, (*c*, *ctc*)), (1, (*g*, *gcc*)), (2, (*c*, *ctc*)), (2, (*g*, *gcc*)), (3, (*c*, *ctc*)), (3, (*g*, *gcc*)), (4, (*c*, *ctc*)), (4, (*g*, *gcc*))
$\hat{\mathcal{P}}_{54}$	(1, (*c*, *gag*)), (1, (*g*, *agc*)), (2, (*c*, *gag*)), (2, (*g*, *agc*)), (3, (*c*, *gag*)), (3, (*g*, *agc*)), (4, (*c*, *ccc*)), (4, (*c*, *gag*)), (4, (*g*, *agc*))
$\hat{\mathcal{P}}_{55}$	(1, (*c*, *gat*)), (1, (*g*, *aat*)), (1, (*t*, *gat*)), (2, (*c*, *gat*)), (2, (*g*, *aat*)), (2, (*t*, *gat*)), (3, (*c*, *gat*)), (3, (*g*, *aat*)), (3, (*t*, *gat*)), (4, (*c*, *gat*)), (4, (*g*, *aat*)), (4, (*t*, *gat*))
$\hat{\mathcal{P}}_{56}$	(1, (*c*, *gca*)), (1, (*c*, *gta*)), (1, (*g*, *acg*)), (2, (*c*, *gca*)), (2, (*c*, *gta*)), (2, (*g*, *acg*)), (3, (*c*, *gca*)), (3, (*c*, *gta*)), (3, (*g*, *acg*)), (4, (*c*, *gca*)), (4, (*c*, *gta*)), (4, (*g*, *acg*))
$\hat{\mathcal{P}}_{57}$	(1, (*c*, *gct*)), (1, (*c*, *ggt*)), (1, (*g*, *gtg*)), (1, (*g*, *tat*)), (2, (*c*, *gct*)), (2, (*c*, *ggt*)), (2, (*g*, *gtg*)), (2, (*g*, *tat*)), (3, (*c*, *gct*)), (3, (*c*, *ggt*)), (3, (*g*, *gtg*)), (3, (*g*, *tat*)), (4, (*c*, *gct*)), (4, (*c*, *ggt*)), (4, (*g*, *gtg*)), (4, (*g*, *tat*))
$\hat{\mathcal{P}}_{58}$	(1, (*c*, *gga*)), (1, (*g*, *cag*)), (1, (*g*, *gga*)), (2, (*c*, *gga*)), (2, (*g*, *cag*)), (2, (*g*, *gga*)), (3, (*c*, *gga*)), (3, (*g*, *cag*)), (3, (*g*, *gga*)), (4, (*c*, *gga*)), (4, (*g*, *cag*)), (4, (*g*, *gga*))
$\hat{\mathcal{P}}_{59}$	(1, (*c*, *gtg*)), (1, (*t*, *cta*)), (2, (*c*, *gtg*)), (2, (*t*, *cta*)), (3, (*c*, *gtg*)), (3, (*t*, *cta*)), (4, (*c*, *gtg*)), (4, (*t*, *cta*))
$\hat{\mathcal{P}}_{60}$	(1, (*c*, *taa*)), (1, (*c*, *tag*)), (1, (*g*, *ctc*)), (1, (*g*, *tca*)), (2, (*c*, *taa*)), (2, (*c*, *tag*)), (2, (*g*, *tca*)), (3, (*c*, *taa*)), (3, (*c*, *tag*)), (3, (*g*, *ctc*)), (3, (*g*, *tca*)), (4, (*c*, *taa*)), (4, (*c*, *tag*)), (4, (*g*, *ctc*)), (4, (*g*, *tca*))

Table 14 (continued)

Part	States
$\hat{\mathcal{P}}_{61}$	$(1, (c, tac)), (1, (c, ttc)), (1, (t, aac)), (1, (t, acc)), (2, (a, tcc)), (2, (c, tac)), (2, (c, ttc)), (2, (t, aac)),$ $(3, (a, tcc)), (3, (c, tac)), (3, (c, ttc)), (3, (t, aac)), (4, (a, tcc)), (4, (c, tac)), (4, (c, ttc)), (4, (t, aac))$
$\hat{\mathcal{P}}_{62}$	$(1, (c, tca)), (1, (g, aca)), (1, (g, cca)), (1, (t, atg)), (2, (c, tca)), (2, (g, aca)), (2, (g, cca)), (2, (t, atg)),$ $(3, (c, tca)), (3, (g, aca)), (3, (g, cca)), (3, (t, atg)), (4, (c, tca)), (4, (g, aca)), (4, (g, cca)), (4, (t, atg))$

Table 15 Composition of parts from $\hat{\mathcal{P}}_{63}$ to $\hat{\mathcal{P}}_{82}$. On the right are the states (by part) $(i, (z, s))$ with $i = 1, ..., 4$ and $(z, s) \in \Lambda(A) = A \times A^o$, with $o = 3$ and $G = 9$ as given by Eq. (2). $i = 1$(B.1.1.7 variant), $i = 2$(B.1.351 variant), $i = 3$(B.1.617.2 variant) and $i = 4$(P.1 variant)

Part	States
$\hat{\mathcal{P}}_{63}$	$(1, (c, tcg)), (1, (g, cga)), (1, (g, tgc)), (2, (c, tcg)), (2, (g, cga)), (2, (g, tgc)), (3, (c, tcg)), (3, (g, cga)),$ $(3, (g, tgc)), (4, (c, tcg)), (4, (g, cga)), (4, (g, tgc))$
$\hat{\mathcal{P}}_{64}$	$(1, (c, tgc)), (1, (t, tgc)), (2, (c, tgc)), (2, (t, tgc)), (3, (c, tgc)), (3, (t, tgc)), (4, (c, tgc)), (4, (t, tgc))$
$\hat{\mathcal{P}}_{65}$	$(1, (c, ttg)), (1, (g, ctg)), (1, (t, ctg)), (2, (c, cgg)), (2, (c, ttg)), (2, (g, ctg)), (2, (t, ccg)), (2, (t, ctg)),$ $(3, (c, cgg)), (3, (c, ttg)), (3, (g, ctg)), (3, (t, ctg)), (4, (c, cgg)), (4, (c, ttg)), (4, (g, ctg)), (4, (t, ccg)),$ $(4, (t, ctg))$
$\hat{\mathcal{P}}_{66}$	$(1, (g, ccc)), (2, (g, ccc)), (3, (g, ccc)), (4, (g, ccc))$
$\hat{\mathcal{P}}_{67}$	$(1, (g, ccg)), (1, (t, acg)), (2, (g, ccg)), (2, (t, acg)), (3, (g, ccg)), (3, (t, acg)), (4, (g, ccg)), (4, (t, acg))$
$\hat{\mathcal{P}}_{68}$	$(1, (g, cgc)), (2, (g, cgc)), (3, (g, cgc)), (4, (g, cgc))$
$\hat{\mathcal{P}}_{69}$	$(1, (g, cgt)), (2, (g, cgt)), (3, (g, cgt)), (4, (g, cgt))$
$\hat{\mathcal{P}}_{70}$	$(1, (g, gat)), (1, (g, gct)), (1, (t, gct)), (2, (g, gat)), (2, (g, gct)), (2, (t, gct)), (3, (g, gat)), (3, (g, gct)),$ $(3, (t, gct)), (4, (g, gat)), (4, (g, gct)), (4, (t, gct))$
$\hat{\mathcal{P}}_{71}$	$(1, (g, ggg)), (1, (g, tcg)), (1, (t, ccg)), (1, (t, tta)), (2, (g, ggg)), (2, (g, tcg)), (2, (t, tta)), (3, (g, ggg)),$ $(3, (g, tcg)), (3, (t, tta)), (4, (g, ggg)), (4, (g, tcg)), (4, (t, tta))$
$\hat{\mathcal{P}}_{72}$	$(1, (g, ggt)), (1, (t, cgt)), (2, (g, ggt)), (2, (t, cgt)), (3, (g, ggt)), (3, (t, cgt)), (4, (g, ggt)), (4, (t, cgt))$
$\hat{\mathcal{P}}_{73}$	$(1, (g, gta)), (2, (g, gta)), (3, (g, gta)), (4, (g, gta))$
$\hat{\mathcal{P}}_{74}$	$(1, (g, tac)), (2, (g, tac)), (3, (g, tac)), (4, (g, tac))$
$\hat{\mathcal{P}}_{75}$	$(1, (t, aag)), (1, (t, gaa)), (2, (t, aag)), (2, (t, gaa)), (3, (t, aag)), (3, (t, gaa)), (4, (t, aag)), (4, (t, gaa))$
$\hat{\mathcal{P}}_{76}$	$(1, (t, agc)), (1, (t, ggc)), (1, (t, gtc)), (2, (t, agc)), (2, (t, ggc)), (2, (t, gtc)), (3, (t, agc)), (3, (t, ggc)),$ $(3, (t, gtc)), (4, (t, agc)), (4, (t, ggc)), (4, (t, gtc))$
$\hat{\mathcal{P}}_{77}$	$(1, (t, att)), (1, (t, ttt)), (2, (t, att)), (2, (t, ttt)), (3, (t, att)), (3, (t, ttt)), (4, (t, att)), (4, (t, ttt))$
$\hat{\mathcal{P}}_{78}$	$(1, (t, ccc)), (1, (t, gcg)), (2, (t, ccc)), (2, (t, gcg)), (3, (t, ccc)), (3, (t, gcg)), (4, (t, ccc)), (4, (t, gcg))$
$\hat{\mathcal{P}}_{79}$	$(1, (t, ggt)), (2, (t, ggt)), (3, (t, ggt)), (4, (t, ggt))$
$\hat{\mathcal{P}}_{80}$	$(1, (t, tat)), (2, (t, tat)), (3, (t, tat)), (4, (t, tat))$
$\hat{\mathcal{P}}_{81}$	$(1, (t, tcc)), (2, (t, tcc)), (3, (t, tcc)), (4, (t, tcc))$
$\hat{\mathcal{P}}_{82}$	$(1, (t, tct)), (1, (t, tgt)), (2, (t, tct)), (2, (t, tgt)), (3, (t, tct)), (3, (t, tgt)), (4, (t, tct)), (4, (t, tgt))$

Joint partition related to Sect. 3.3

This part of the appendix shows the composition of the entire joint partition $\hat{\mathbb{P}}$ on $\{1, 2, 3, 4\} \times \Lambda(A)$, corresponding to the multiple partition model, estimated in Sect. 3.3 from the selected sequences.

Tables 12, 13, 14, and 15, contain the elements of each of the 82 parts composing the full joint partition $\hat{\mathbb{P}}$.

Acknowledgements The authors wish to thank the two referees and editors for their many helpful comments and suggestions on an earlier draft of this paper.

Declarations

Funding Author G. H. Tasca has received partial financial support, provided by Coordenação de AperfeiÃ§oamento de Pessoal de Nível Superior (CAPES) with a fellowship from the Ph.D. Program in Statistics - University of Campinas.

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Availability of data and material All genome sequences data were obtained at the https://www.gisaid.org/ and https://www.ncbi.nlm.nih.gov/ repositories.

Code availability Not applicable.

References

Buhlmann P, Wyner AJ (1999) Variable length Markov chains. Ann Stat 27(2):480–513. https://doi.org/10.1214/aos/1018031204

Christley S, Lu Y, Li C, Xie X (2009) Human genomes as email attachments. Bioinformatics 25(2):274–275. https://doi.org/10.1093/bioinformatics/btn582

Cordeiro MTA, García JE, González-López VA, Mercado Londoño SL (2020) Partition Markov model for multiple processes. Math Methods Appl Sci 43(13):7677–7691. https://doi.org/10.1002/mma.6079

Csiszár I, Talata Z (2006) Context tree estimation for not necessarily finite memory processes, via BIC and MDL. IEEE Trans Inf Theory 52(3):1007–1016. https://doi.org/10.1109/TIT.2005.864431

Deng X, Garcia-Knight MA, Khalid MM, Servellita V, Wang C, Morris MK, Sotomayor-González A, Glasner DR, Reyes KR, Gliwa AS, Reddy NP, Sanchez San Martin C, Federman S, Cheng J, Balcerek J, Taylor J, Streithorst JA, Miller S, Sreekumar B, Chen P-Y, Schulze-Gahmen U, Taha TY, Hayashi JM, Simoneau CR, Renuka Kumar G, McMahon S, Lidsky PV, Xiao Y, Hemarajata P, Green NM, Espinosa A, Kath C, Haw M, Bell J, Hacker JK, Hanson C, Wadford DA, Anaya C, Ferguson D, Frankino PA, Shivram H, Lareau LF, Wyman SK, Ott M, Andino R, Chiu CY (2021) Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant. Cell 184(13):3426-3437.e8. https://doi.org/10.1016/j.cell.2021.04.025

Fernández M, García JE, Gholizadeh R, González-López VA (2020) Sample selection procedure in daily trading volume processes. Math Methods Appl Sci 43(13):7537–7549. https://doi.org/10.1002/mma.5705

García JE, González-López VA, Tasca GH (2022) A stochastic inspection about genetic variants of COVID-19 circulating in Brazil during 2020. In: AIP Conference Proceedings (vol 2425, no 1, p 230002). AIP Publishing LLC. https://doi.org/10.1063/5.0081337

García JE, González-López VA (2017) Consistent estimation of partition Markov models. Entropy 19(4):160. https://doi.org/10.3390/e19040160

García JE, Gholizadeh R, González-López VA (2018) A BIC-based consistent metric between Markovian processes. Appl Stoch Models Bus Ind 34(6):868–878. https://doi.org/10.1002/asmb.2346

García JE, González-López VA, Tasca GH (2020) Partition Markov model for Covid-19 virus. 4open 3:13. https://doi.org/10.1051/fopen/2020013

Hirotsu Y, Omata M (2021) Discovery of a SARS-CoV-2 variant from the P.1 lineage harboring K417T/E484K/N501Y mutations in Kofu, Japan. J Infect 82(6):276–316. https://doi.org/10.1016/j.jinf.2021.03.013

Hoffmann M, Arora P, Groß R, Seidel A, Hörnich B, Hahn A, Krüger N, Graichen L, Hofmann-Winkler H, Kempf A, Winkler MS, Schulz S, Jäck HM, Jahrsdörfer B, Schrezenmeier H, Müller M, Kleger A, Münch J, Pöhlmann S (2021) SARS-CoV-2 variants B.1.351 and B.1.1.248: escape from therapeutic antibodies and antibodies induced by infection and vaccination. Cell 184(9):2384–2393. https://doi.org/10.1101/2021.02.11.430787

Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley, Hoboken

Nonaka CK, Franco MM, Gräf T, de Lorenzo Barcia CA, de Ávila Mendonça RN, De Sousa KAF, Neiva LMC, Fosenca V, Mendes AVA, de Aguiar RS, Giovanetti M, de Freitas Souza BS (2021) Genomic evidence of SARS-CoV-2 reinfection involving E484K spike mutation, Brazil. Emerg Infect Dis 27(5):1522. https://doi.org/10.3201/eid2705.210191

Rissanen J (1983) A universal data compression system. Trans Inf Theory 29(5):656–664. https://doi.org/10.1109/TIT.1983.1056741

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464. https://doi.org/10.1214/aos/1176344136

Taylor L (2021) Covid-19: How the Brazil variant took hold of South America. BMJ. https://doi.org/10.1136/bmj.n1227

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.