4open 2020, **3**, 13

https://doi.org/10.1051/fopen/2020013

4 Open

Available online at:
www.4open-sciences.org

RESEARCH ARTICLE OPEN ACCESS

Partition Markov Model for Covid-19 Virus

Jesús Enrique García, Verónica Andrea González-López, and Gustavo Henrique Tasca*

Department of Statistics, University of Campinas, Sergio Buarque de Holanda, 651, 13083-859 Campinas, S.P., Brazil

Received 12 March 2020, Accepted 13 August 2020

Abstract – In this paper, we investigate a specific structure within the theoretical framework of *Partition Markov Models* (PMM) [see García Jesús and González-López, Entropy 19, 160 (2017)]. The structure of interest lies in the formulation of the underlying *partition*, which defines the process, in which, in addition to a finite memory o associated with the process, a parameter G is introduced, allowing an extra dependence on the past complementing the dependence given by the usual memory o. We show, by simulations, how algorithms designed for the classic version of the PMM can have difficulties in recovering the structure investigated here. This specific structure is efficient for modeling a complete genome sequence, coming from the newly decoded Coronavirus Covid-19 in humans [see Wu et al., Nature 579, 265–269 (2020)]. The sequence profile is represented by 13 units (parts of the state space's partition), for each of the 13 units, their respective transition probabilities are computed for any element of the genetic alphabet. Also, the structure proposed here allows us to develop a comparison study with other genomic sequences of Coronavirus, collected in the last 25 years, through which we conclude that Covid-19 is shown next to SARS-like Coronaviruses (SL-CoVs) from bats specimens in Zhoushan [see Hu et al., Emerg Microb Infect 7, 1–10 (2018)].

Keywords: Bayesian information criterion, Partition Markov Models, Metric between Markov processes

Introduction

A Partition Markov Model (PMM) is a model established in a discrete stochastic process on a finite alphabet, with finite order, see [1]. PMM generalizes other models, including Variable Length Markov chains and Markov chains with finite order. A PMM identifies a partition in the state space, bringing together in each part of the partition, the states that share the same transition probabilities for all the elements in the alphabet. All the states of a part are used to compute the transition probabilities, allowing us to use several states (all those included in the part) to estimate a unique set of transition probabilities. By construction, this is a parsimonious model, since it reduces the number of probabilities to estimate by identifying equivalent states (found in the same part). PMM models have already shown sufficient flexibility in the field of genomic structure modeling, for several purposes, such as determining similarities between Zika's genomic sequences and modeling the genomic Zika's profile [2]. This family of models has also been used to define the genomic profile of the Epstein–Barr virus [3]. The consistent estimation of a PMM [1], is achieved by the Bayesian Information Criterion, BIC, which has led to the definition of a BIC-based metric, see also [4]. The metric has allowed the use of the dynamics of the PMM models for other open problems; it has been efficiently applied in subjects such as (1) the comparison between Dengue's genomic sequences of different origins [5] and (2) to compare genomic sequences of Burkitt lymphoma/leukemia [6]. Since a PMM is defined by a partition on the state space of the process, we understand that the partition's structure could be the key to modeling certain phenomena. Then, in this paper, we investigate specific impositions on the partition of a PMM, with the purpose of improving the modeling of genomic sequences. Given a memory o, the states are sequences of o elements of the alphabet. The state space is the set of states, and the partition of the PMM is a partition of the state space. Each state is a configuration that occurs in a consecutive interval of time, so in order to define the next element of the process (the transition) is necessary to observe a past of size o. In addition to a finite memory o, this paper introduces a parameter G. G allows considering the dependence on previous events in the past of the process realization, that are not accounted for the memory o. The PMM model thus specified could be used to achieve more intricate dependence structures allowing the representation of genomic sequences and, that is the objective of this paper, to verify if, in fact, we can achieve

*Corresponding author: tasca_gustavo@hotmail.com

a finer model for the genomic structure of a complete DNA sequence, of the outbreak of a novel Coronavirus (Covid-19), collected in Wuhan of Hubei province, China. In addition to tracing the genomic profile of Covid-19, we also want to compare genomic sequences that according to the literature could point to the origins of the sequence used in this article, considered one of the first records of the virus.

This article is organized as follows. Section Theoretical Background addresses the theoretical foundations of Partition Markov Models, the estimation process, and the specific case of PMM that we investigate in this article. Section Covid-19 DNA Model describes a real problem associated with the identification of the profile of a new Coronavirus named Covid-19. This section describes the data and how the specific case of PMM shows an improved performance to describe the stochastic behavior of the new virus. The conclusions and considerations are given in Section Conclusion.

Theoretical background

In this section, we present the notation as well as the fomalization of the model on which we developed our discussion, the *Partition Markov Model*. We also show how the aforementioned model can be consistently estimated. Let $(Z_t)_{t\geq 1}$ be a discrete time Markov chain of order o on a finite alphabet A, such that $o < \infty$. Let us call $\mathcal{S} = A^o$ the state space and denote the string $a_m\ a_{m+1},\ \ldots,\ a_n$ by a_m^n where $a_i \in A$, $m \leq i \leq n$. For each $a \in A$ and $s \in \mathcal{S}$ define the transition probability $P(a|s) = \text{Prob}\,(Z_t = a|Z_{t-o}^{t-1} = s)$. In a given sample z_1^n, coming from the stochastic process, the number of occurrences of s is denoted by $N_n(s)$ and the number of occurrences of s followed by a is denoted by $N_n(s,\ a)$. In this way, $\frac{N_n(s,a)}{N_n(s)}$ is the maximum likelihood estimator of $P(a|s)$. The Partition Markov Model introduced in the next definition is designated to obtain a parsimonious model for a Markov process with finite memory on a finite alphabet. This model proposes the identification of states in the state space in units called parts (of a partition), the parts are composed by states which have in common their transition probabilities.

Definition 2.1. *Let $(Z_t)_{t\geq 1}$ be a discrete time Markov chain of order o on a finite alphabet A, with state space $\mathcal{S} = A^o$,*

(i) *$s,\ r \in \mathcal{S}$ are equivalent if $P(a|s) = P(a|r)\ \forall a \in A$.*

(ii) *$(Z_t)_{t\geq 1}$ is a Markov chain with partition $\mathcal{L}= \{L_1,\ L_2,\ \ldots,\ L_{|\mathcal{L}|}\}$ if this partition is the one defined by the equivalence relationship introduced by item i.*

The model given by Definition 2.1 was introduced in [1]. The parameters to be estimated are (a) the partition $\mathcal{L}$, (b) the transition probabilities of each part L to any element of A, $P(\cdot|L)$ which is $P(\cdot|L) = P(\cdot|s)$, $\forall s \in L$. We note that the partition of $\mathcal{S}$ that responds to item (ii) of Definition 2.1 is minimal in relation to the number of parts $|\mathcal{L}|$. Given a sample of $(Z_t)_{t\geq 1}$, z_1^n, according to [1] the partition can be consistently estimated by means of $d_{\mathcal{L}}$ given by Definition 2.2.

Definition 2.2. *Let $(Z_t)_{t\geq 1}$ be a discrete time Markov chain of order o on a finite alphabet A, with state space $\mathcal{S} = A^o$, z_1^n a sample of the process and let $\mathcal{L} = \{L_1,\ L_2,\ \ldots,\ L_{|\mathcal{L}|}\}$ be a partition of $\mathcal{S}$ such that for all $s,\ r \in L_l$, $P(\cdot/s) = P(\cdot/r)$ for each $l = 1,\ 2,\ \ldots,\ |\mathcal{L}|$,*

$$d_{\mathcal{L}}(i,j) = \frac{\alpha}{(|A|-1)\ln(n)} \sum_{a\in A} \left\{ \sum_{l=i,j} N_n(L_l,a) \ln\left(\frac{N_n(L_l,a)}{N_n(L_l)}\right) - N_n(L_{ij},a) \ln\left(\frac{N_n(L_{ij},a)}{N_n(L_{ij})}\right) \right\},$$

with $N_n(L_i) = \sum_{s\in L_i} N_n(s)$, $N_n(L_i,a) = \sum_{s\in L_i} N_n(s,a), N_n(L_{ij}) = N_n(L_i) + N_n(L_j), N_n(L_{ij},a) = N_n(L_i,a) + N_n(L_j,a)$; for $a \in A$, $L_i,\ L_j \in \mathcal{L}, \alpha$, a real and positive value.

The estimation of Partition Markov Models can be carried out via algorithms such as the one introduced in [7], which uses $d_{\mathcal{L}}$. Note that $d_{\mathcal{L}}$ is a metric designed to build a structure in the state space, identifying equivalent states, it is applied (see the algorithm of [7]) for example in an initial set consisting of the entire state space $\mathcal{S}$, and whenever $d_{\mathcal{L}}(i,\ j) < 1$ the elements L_i and L_j must be in the same part (see properties of $d_{\mathcal{L}}$ in [1]). The metric $d_{\mathcal{L}}$ is derived from the Bayesian Information Criterion (BIC), as proved in [1], and the BIC indicates the junction of two elements in the same part of the partition, if and only if $d_{\mathcal{L}} < 1$. For each part L of $\hat{\mathcal{L}}$ the transition probability is estimated by $\hat{P}(a|L) = \frac{N_n(L,a)}{N_n(L)}$. Note that all equivalent states are used to estimate each probability. An economy is produced in the total number of probabilities to be estimated, since the identification of the partition produces a reduction of the number of probabilities to be estimated and each probability can be better estimated since the occurrences of several states are used for the estimation of each probability, now related to the part of the partition $P(\cdot|L)$.

In the next subsection, we discuss a specific structure of the partition $\mathcal{L}$ and how it might make sense in practice. We also discuss the impact of such specificity on estimation algorithms, such as the one exposed in [7].

A further memory for the process

When fitting a PMM, the state space is restricted by the maximum value for the memory o allowed by the sample size and, if the alphabet is A, the space where the partition is arranged is build in $\mathcal{S} = A^o$. This means that to define the next step of the process, it is enough to know the past of size o, but it could happen that in addition to a past of size o, the process depends on a more distant jump, say G, with $G > o$. It is natural to think then in defining the state space as A^G, where we would again have the structure of a Partition Markov Model, but with a redundancy of information, since the values between times $t - G + 1$ and $t - o - 1$ are not relevant for the future, see Figure 1 to illustrate the idea (the zigzag part is irrelevant). Let see the formalization of the situation, suppose that there is a value $G > o$, such that the transition probability to $Z_t = a \in A$ from $Z_{t-G}^{t-1} = z \dots s$, $z \in A$, $s \in A^o$, where "..." is any concatenation of elements of A of size $G - o - 1$, is given by,

$$\text{Prob}\,(Z_t = a | Z_{t-G}^{t-1} = z \dots s) = \text{Prob}\,(Z_t = a | Z_{t-G} = z, Z_{t-o}^{t-1} = s),\ z \in A,\ s \in A^o. \tag{1}$$

Then, the space is given by $A \times A^o$, where A records all possibilities for z and A^o records all possibilities for s, (z, s) of equation (1).

The kind of process on which we want to identify the partition of the state space is given to follow.

Definition 2.3. *A G-Markov Model $(Z_t)_{t\geq 1}$ is a discrete time Markov chain on a finite alphabet A, with state space $\mathcal{W} = A \times A^o$, where $o < \infty$, transition probabilities following equation (1), for an adequate and finite G such that $G > o$.*

Given a sample z_1^n, the number of ocurrences of $(z, s) \in A \times A^o$ in the sample is $N_n(z, s) = |\{t : G < t \leq n, z_{t-G} = z, z_{t-o}^{t-1} = s\}|$ and the ocurrences of $(z, s) \in A \times A^o$ followed by $a \in A$ is $N_n((z, s), a) = |\{t : G < t \leq n, z_{t-G} = z, z_{t-o}^{t-1} = s, z_t = a\}|$.

Remark 2.1

Itens (i) and (ii) of Definition 2.1 allow defining the partition of $(Z_t)_{t\geq 1}$, say $\mathcal{I}$ (partition of $\mathcal{W}$ of Definition 2.3). We can also adapt the metric of Definition 2.2 to this situation, to do that it is enough to change $\mathcal{S}$ by $\mathcal{W}$, defining for each part $\mathcal{I}$ of the partition $\mathcal{I}$ of $\mathcal{W}$, $N_n(I) = \sum_{(z,s)\in I} N_n(z, s)$ and $N_n(I, a) = \sum_{(z,s)\in I} N_n((z, s), a)$, for $a \in A$. Denote by $d_{\mathcal{I}}$ the metric on the state space $\mathcal{W}$. So, to estimate $\mathcal{I}$ we can use the algorithm introduced in [7].

Given a sample z_1^n of the process $(Z_t)_{t\geq 1}$, denote by $P(a|(z, s))$ the transition probability given by equation (1), then, the likelihood of the sample $P(z_1^n) = \text{Prob}\,(Z_1^n = z_1^n)$ is,

$$P(z_1^n) = P(z_1^G) \prod_{a\in A, I\in I} P(a|I)^{N_n(I,a)}, \tag{2}$$

which is the same expression of equation (2) of [1]. Then, the procedure is to compute the BIC for each model $\mathcal{I}$, which is given by equation (3),

$$\text{BIC}\,(z_1^n, \mathcal{I}) = \ln\left(\prod_{a\in A, I\in\mathcal{I}} \left(\frac{N_n(I, a)}{N_n(I)}\right)^{N_n(I,a)}\right) - \frac{(|A| - 1)|\mathcal{I}|\ln(n)}{\alpha}. \tag{3}$$

As the BIC definition itself shows, the models are characterized (Eq. (3)) by the logarithm of the maximum likelihood, $\ln\left(\prod_{a\in A, I\in\mathcal{I}} \left(\frac{N_n(I,a)}{N_n(I)}\right)^{N_n(I,a)}\right)$, penalized by the number of parameters to be estimated, $(|A| - 1)|\mathcal{I}|$, properly scaled by the size of the data set (by the term $\ln(n)/\alpha$). Thus, the model indicated by the BIC is the one with the highest BIC value that corresponds to the most plausible, taking into account the complexity of the model (number of parameters). The criterion is derived in [8], using $\alpha = 2$, and it corresponds to the maximization of a posterior distribution assuming a non-informative prior distribution on the dimension of the parametric space. We note that the BIC continues to be valid, replacing the constant 2 with any positive constant α, as given in equation (3).

By means of the maximization of equation (3) the partition can be estimated, obtaining $\hat{\mathcal{I}}$ as equation (4),

Figure 1. Scheme of the past necessary to determine the state of the process at time t, according to equation (1). In zigzag the irrelevant period with limits on top of the scheme $[t - G + 1,\ t - o - 1]$.

$$\hat{\mathcal{I}} = \text{argmax}_{\mathcal{I}\in\mathcal{P}}\{\text{BIC}\,(z_1^n, \mathcal{I})\}, \tag{4}$$

where $\mathcal{P}$ is the set of all the partitions of $\mathcal{W}$.

As the set $\mathcal{P}$ can be huge, to obtain $\hat{\mathcal{I}}$ it is necessary to use the metric $d_{\mathcal{I}}$ with the algorithm introduced by [7].

Remark 2.2

Note that under the Definition 2.3, for the construction of equation (3) and subsequent derivation of the maximum given in equation (4), the parameters G and o are previously set.

The BIC criterion shows great advantages, and therefore, its use is recommended. We quote some of its qualities, (i) it is a consistent method for the estimation of models given by Definition 2.3 (Theorem 3 – [1]). Already particular cases of Definition 2.3 anticipated the consistency of the BIC for the estimation, see, for example [9], in the framework of variable-length Markov chains. (ii) the BIC allows creating a metric like the one detailed in Remark 2.1 (Theorem 2, corollary 2 – [1]), which facilitates the implementation of algorithms [7]. The second property imposes the preference of the BIC against other criteria such as the Krichevsky–Trofimov (KT) [9].

The structure that we want to investigate in this paper is about the form of the partition, which responds to the rule illustrated in Figure 1. In the next example, we present a case.

Example 2.1

Consider a G-Markov model $(Z_t)_{t\geq 1}$ (Definition 2.3) with $A = \{a, b, c\}$, $o = 1$, $G = 4$, partition $\mathcal{I} = \{I_1, I_2, I_3\}$ and transition probabilities given by Table 1. For instance, since $P(a|I_1) = 0.2$ we have that $P(Z_t = a|Z_{t-4}^{t-1} = a\bigstar * a) = 0.2$, $\forall\bigstar \in A$ and $\forall * \in A$. Then, the values $\bigstar$ and $*$ at positions $t-3$ and $t-2$ respectively (between the times $t-4$ and $t-1$) are irrelevant for the state at time t.

In the following two simulations we see the impact of the structure of a G Markov model – Definition 2.3 – when we ignore it and apply the algorithm of [7] with the help of $d_{\mathcal{L}}$, assuming only the usual PMM structure – Definition 2.1.

Example 2.2

We apply the algorithm of [7] in a simulated data from the law given by Table 1, the algorithm is applied in two settings (i) using $o = 4$, initial set $\{a, b, c\}^4$ and $d_{\mathcal{L}}$ as given by Definition 2.2 and, (ii) using $o = 1$, $G = 4$, initial set $\{a, b, c\} \times \{a, b, c\}$ and $d_{\mathcal{I}}$ with the modifications imposed by Remark 2.1. This means that in (i) we fit a Partition Markov Model without a parameter G, as usual and in (ii) a Partitition Markov Model with a G (Definition 2.3). With a sample size $n = 5 \times 10^4$, we obtain by (ii) the original partition (Tab. 1). By (i) the partition obtained is given in Table 2. Note that under the setting (i) each element of the state space is a concatenation of $o = 4$ consecutive elements of A, for example $accb$, and this element under the setting (ii) is denoted by (a, b) where the memory $o = 1$ is related to the element b and $G = 4$ is related to the element a, being irrelevant the central elements cc. Also observe that there is a relationship between some parts of Tables 1 and 2, $L_1 = I_2$ and $L_5 = I_3$, but the part I_1 is distributed in the parts: L_2, L_3 and L_4. That is to say that when adjusting via (i) a confusion of the original structure (Tab. 1) is generated.

To visualize the behavior of the settings (i) and (ii) when increasing the sample size, for each sample size $n = 5 \times 10^4$, 10^6, 5×10^6, 10^7 we perform 100 simulations of the model – Table 1. We record the performance of the settings (i), (ii) in

Table 1. Partition $\mathcal{I}$ and transition probability to each element in the alphabet A.

Part	Elements	$P(a\|I)$	$P(b\|I)$
I_1	(a, a), (a, b), (a, c), (b, a), (b, c), (c, c)	0.2	0.3
I_2	(b, b), (c, b)	0.4	0.3
I_3	(c, a)	0.4	0.1

Table 2. Estimated partition for the state space $\{a, b, c\}^4$ – procedure (i). The elements of L_1 have the format $b \ldots b$ or $c \ldots b$ and the elements of L_5 have the format $c \ldots a$.

Part	Elements
L_1	*bbbb*, *cbbb*, *babb*, *cabb*, *bcbb*, *ccbb*, *bbab*, *cbab*, *baab*, *caab*, *bcab*, *ccab*, *bbcb*, *cbcb*, *bacb*, *cacb*, *bccb*, *cccb*
L_2	*abbb*, *acbb*, *bcba*, *bbca*, *baca*, *aaca*, *acca*, *abbc*, *babc*, *aabc*, *baac*, *aaac*, *bcac*, *bacc*, *aacc*
L_3	*aabb*, *aaab*, *abcb*, *aacb*, *accb*, *bbba*, *abba*, *acba*, *abaa*, *baaa*, *bcaa*, *abca*, *bcca*, *bbbc*, *cbbc*, *bcbc*, *acbc*, *bbac*, *abac*, *caac*, *ccac*, *cacc*, *cccc*
L_4	*abab*, *acab*, *baba*, *aaba*, *bbaa*, *aaaa*, *acaa*, *cabc*, *ccbc*, *cbac*, *acac*, *bbcc*, *abcc*, *cbcc*, *bccc*, *accc*
L_5	*cbba*, *caba*, *ccba*, *cbaa*, *caaa*, *ccaa*, *cbca*, *caca*, *ccca*

Table 3. Number of parts of the partitions estimated for settings (i) and (ii) in 100 simulations of size n each.

(i) Number of parts in $\mathcal{S} = \{a, b, c\}^4$					(ii) Number of parts in $\mathcal{W} = \{a, b, c\} \times \{a, b, c\}$		
n	3 parts	4 parts	5 parts	6 parts	n	3 parts	4 parts
5×10^4	0	58	40	2	5×10^4	99	1
10^6	6	84	10	0	10^6	100	0
5×10^6	26	72	2	0	5×10^6	100	0
10^7	32	66	2	0	10^7	100	0

Table 4. Partition $\mathcal{I}$ and transition probability of each part I_i, $i = 1, 2$ to each element in the alphabet A.

Part	Elements	$P(a\|I)$	$P(b\|I)$
I_1	a, b	0.2	0.3
I_2	c	0.4	0.3

recovering the appropriate number of parts of the partition (which is 3). The results are shown in Table 3. We see how (ii) has a more efficient behavior, showing that a usual procedure, such as (i) can show difficulties in recovering the structure given by Definition 2.3. Note also that all the parts recovered by (ii) in the last 3 sample sizes are the original ones – see Table 1.

In practical terms, when modeling with real data, a specific sample size is available, say n. In general, the memory of the process that can be used in the model depends on n as well as the cardinal $|A|$ of the process alphabet. Usually, the memory must be less than $\log_{|A|}(n)$, in the next example, we see the effect of this condition on G.

Example 2.3

Consider a G-Markov model $(Z_t)_{t\geq 1}$ (Definition 2.3) with $A = \{a, b, c\}$, $o = 0$ and $G = 10$, with partition $\mathcal{I} = \{I_1, I_2\}$ and transition probabilities given by Table 4. We perform simulations of the process following Table 4, with $n = 2000$. The algorithm of [7] is applied in two settings (i) using $o = 4 < \lfloor \log_{|A|}(n) \rfloor - 1 = 5$, initial set $\{a, b, c\}^4$ and $d_{\mathcal{L}}$ as given by Definition 2.2 and, (ii) using $o = 0$, $G = 10$, initial set $\{a, b, c\}$ and $d_{\mathcal{I}}$ with the modifications imposed by Remark 2.1. In other words, case (i) reflects the conditions that are generally applied to proceed with the adjustment and determination of the process memory.

In Table 5 we show the resulting partition of (i). We see that it is not possible to recover the original structure given in Table 4. As expected, (ii) recovers the structure given by Table 4. These results reflect the insufficiency for $n = 2000$ to reach a memory that encompasses $G = 10$, which is the period that the process needs to determine the choice of the next step.

Section Covid-19 DNA Model shows how the model stands out when it comes to representing the genome of Covid-19.

Covid-19 DNA model

In this section, we investigate the stochastic behavior of a complete DNA sequence of the outbreak of a novel Coronavirus (Covid-19) associated with a respiratory disease in Wuhan of Hubei province, China. The sequence was extracted from a patient coming from the Wuhan seafood market, a place that was associated with the origin of the outbreak, its accession number is MN908947 (version MN908947.3). The sequence is coming from a 41-year-old man with no history of hepatitis, tuberculosis or diabetes [10]. The patient was admitted and hospitalized in Wuhan Central Hospital on December 26, 2019, 6 days after the onset of illness. He reported fever, chest tightness, cough, pain, and weakness for one

Table 5. Estimated partition for the state space $\{a, b, c\}^4$ – procedure (i).

Part	Element	Cardinal
L_1	*aaaa*, *aaac*, *aabb*, *aacb*, *abac*, *abbc*, *abcc*, *acaa*, *acbb*, *baab*, *baca*, *bbab*, *bbbb*, *bbcb*, *bbcc*, *bcaa*, *bccb*, *cabb*	18
L_2	*aaca*, *abaa*, *baaa*, *babc*, *bacb*, *bcab*, *bcba*, *caaa*, *caab*, *caba*, *caca*, *cbbb*, *cbca*, *ccac*, *ccca*	15
L_3	*abcb*, *baba*, *babb*, *bbaa*, *bbca*, *bcbc*	6
L_4	*abbb*, *abca*, *bbba*, *bcac*, *bcbb*, *cbaa*, *cbba*, *cbcc*, *ccba*	9
L_5	*aaab*, *aaba*, *aabc*, *aacc*, *abab*, *abba*, *acab*, *acac*, *acba*, *acbc*, *acca*, *accb*, *accc*, *baac*, *bacc*, *bbac*, *bbbc*, *bcca*, bccc,*caac*, *cabc*, *cacb*, *cacc*, *cbab*, *cbac*, *cbbc*, *cbcb*, *ccaa*, *ccab*, *ccbb*, *ccbc*, *cccb*, *cccc*	33

Table 6. On top, settings under the perspective (i): memory o, cardinal of partition $|\mathcal{L}|$, BIC and KT values. On bottom, settings under the perspective (ii): parameter G, cardinal of partition $|\mathcal{I}|$, BIC and KT values. In bold letter the 2 best cases.

(i)			
o	$\lvert\mathcal{L}\rvert$	BIC	KT
1	4	−39 965.08	39 957.9
2	7	−39 925.36	39 902.16
3	9	**−39 832.03**	**39 800.01**
4	15	**−39 661.36**	**39 594.58**
(ii) with $o = 3$			
G	$\lvert\mathcal{I}\rvert$	BIC	KT
5	14	−39 670.76	39 609.50
6	15	−39 663.48	39 596.82
7	13	−39 678.24	39 621.91
8	14	−39 690.33	39 628.22
9	13	**−39 639.89**	**39 585.72**
10	12	−39 673.21	39 623.47
11	14	−39 687.28	39 626.58
12	15	**−39 660.44**	**39 592.87**

week. The cardiovascular, abdominal, and neurologic examination was normal; see more details in [10]. The sequence can be obtained from https://www.ncbi.nlm.nih.gov/nuccore/MN908947. We use in this paper the *FASTA* format of MN908947, which is composed by 29 903 bases: a, c, g, t.

For the construction of the model, we must choose a memory o. In a discrete Markov process with a discrete alphabet, the criterion used is $o < \left\lfloor \log_{|A|}(n) \right\rfloor - 1$. So, for the alphabet $A = \{a, c, g, t\}$ with $n = 29\,903$, we have the restriction $o < 6$. Since the bases in a DNA structure are organized in triples, it is recommended $o = 3$. This organization also applies to memory G, it is expected that the values of G multiples of 3 show better performance.

We apply the algorithm of [7] in two settings (i) using $o \in \{1, 2, 3, 4\}$, initial set $\{a, c, g, t\}^o$ and $d_{\mathcal{L}}$ as given by Definition 2.2, (ii) using $o = 3$, $G \in \{5, 6, 7, 8, 9, 10, 11, 12\}$, initial set $\{a, c, g, t\} \times \{a, c, g, t\}^3$ and $d_{\mathcal{I}}$ with the modifications imposed by Remark 2.1. To identify the best model for the sequence, we use the BIC criterion (with $\alpha = 2$), see equation (3). And, therefore, the higher the BIC value, the better the model represents the sequence. To compare the results, we also report the Krichevsky–Trofimov (KT) criterion [9]. According to the KT definition, the smaller the value, the better the model will be for representing the data. According to Table 6, the best models are given by two models in (ii). This shows us the convenience of assuming the existence of an extra parameter G. Note that in the three best cases of (ii) G is a multiple of 3, $G = 9$, 12, 6, which confirms the nature of the genomic organization in triples formed by elements of the alphabet A.

As is usual in DNA sequences [2], the transition probabilities are moderate (in this case ≤ 0.44, see Tab. 8). Note that there is a predilection of the process to choose as the next state a or t. Sequences of other viruses lead to other predilections, see for example [2], in which the Zika process is modeled, revealing a predilection for the states a or g. We observe that under Definition 2.3 and without imposing the partition structure, the total number of parameters to be estimated is unfeasible. For example, with $o = 3$ and $G = 9$ we have $|A \times A^o| \times (|A| - 1) = 768$ parameters to estimate, and when using the strategy given by equation (2) and Remark 2.1, it is necessary to estimate $|\mathcal{I}| \times (|A| - 1) = 39$ parameters. Table 7 shows the composition of each part, for example, part I_1 is composed of 31 elements of type (z, s) where $z \in A$ and $s \in A^o$. The elements are listed from left to right according to how they have been inserted in the part by the algorithm of [7] and following Remark 2.1.

Covid-19 and coronaviruses

In this subsection, we incorporate 11 sequences into the study to compare how distant or close to them the sequence investigated is. It is speculated that the new sequence is the product of mutations of other types of Coronaviruses, and a way to deal with it could be to determine those sequences that are the closest. Although Coronaviruses similar to severe acute respiratory syndrome (SARS, see [11]) have been widely identified in mammals, including bats, since 2005 in China, the exact origin of Coronaviruses infecting humans remains unclear. Therefore, it is necessary to determine the natural reservoir and any intermediate hosts of Coronavirus in its current version (Covid-19). We describe the sequences in Table 9, those are complete sequences of the Coronavirus genome of different types that occurred in the last 25 years.

The metric introduced to follow makes this comparison possible.

Table 7. Part composition for the model – (ii) with $o = 3$ and $G = 9$, see Table 6.

Part	Elements
1	(a, aaa), (a, aca), (a, aga), (a, ata), (a, cca), (a, cga), (a, gca), (a, gta), (a, tca), (c, aag), (c, aca), (c, aga), (c, caa), (c, cta), (c, gga), (c, taa), (c, tag), (c, tta), (g, aga), (g, cag), (g, ctc), (g, tag), (g, tca), (g, tta), (t, aag), (t, aca), (t, aga), (t, cag), (t, cca), (t, cga), (t, tca)
2	(a, atg), (a, cgt), (a, ctg), (a, gag), (a, ttg), (c, act), (c, att), (c, tca), (g, aaa), (g, aca), (g, atg), (g, att), (g, cca), (g, ctt), (t, aaa), (t, atg), (t, cat)
3	(a, caa), (a, ccg), (a, cgc), (a, gga), (c, aaa), (c, ata), (c, cct), (c, gaa), (c, gca), (c, gta), (g, acg), (g, ata), (g, caa), (g, gca), (g, gga), (g, gta), (t, gca), (t, gga), (t, gta)
4	(a, aag), (a, cag), (a, cta), (a, taa), (a, tga), (c, atc), (c, ggc), (c, gtc), (c, tga), (g, aag), (g, agg), (g, taa), (g, tcc), (g, ttc), (t, ata), (t, taa), (t, tag)
5	(a, gaa), (a, gat), (a, ggt), (a, tat), (c, cgt), (c, gat), (g, aat), (g, act), (g, agt), (g, cgt), (g, gaa), (g, gct), (g, ggt), (g, gtt), (g, tgt), (t, cgt), (t, gat), (t, gct)
6	(a, att), (a, cct), (a, gct), (a, gtt), (a, ttt), (c, agg), (c, cag), (c, ctt), (c, ttt), (g, cct), (g, tct), (g, ttt), (t, att), (t, gaa), (t, gag), (t, ggt), (t, ttt)
7	(a, tta), (c, cac), (c, gtg), (c, tgg), (g, cac), (g, cgg), (g, cta), (g, ggg), (g, tcg), (t, caa), (t, cta), (t, tga), (t, tta)
8	(a, acg), (a, act), (a, agg), (a, ggg), (a, tct), (a, tgg), (a, tgt), (c, acg), (c, ggg), (c, tct), (c, tgt), (c, ttg), (g, ccg), (g, ggc), (g, tga), (g, tgg), (t, acg), (t, act), (t, agg), (t, ccc), (t, cct), (t, ctg), (t, ctt), (t, gcg), (t, gtt), (t, tct), (t, tgt)
9	(a, agc), (a, atc), (a, cgg), (a, gcg), (a, gtc), (a, tag), (a, tcc), (a, tgc), (a, ttc), (c, acc), (c, agc), (c, ccg), (c, cga), (c, gcc), (c, tac), (c, tcc), (c, tgc), (g, cga), (g, tgc), (t, aac), (t, acc), (t, agc), (t, atc), (t, cac), (t, ggc), (t, ggg), (t, gtc), (t, tac), (t, tcc), (t, tgc), (t, tgg), (t, ttc)
10	(a, cat), (a, ctt), (a, gtg), (a, tcg), (c, atg), (c, cat), (c, cca), (c, cgg), (c, ctg), (c, gcg), (g, ctg), (g, gag), (g, gcg), (g, ttg), (t, ccg), (t, cgg), (t, gtg), (t, tcg), (t, ttg)
11	(a, aat), (a, agt), (c, aat), (c, agt), (c, gct), (c, ggt), (c, gtt), (c, tat), (g, cat), (g, gat), (g, gtg), (g, tat), (t, aat), (t, agt), (t, tat)
12	(a, acc), (a, gcc), (a, ggc), (c, ccc), (c, cgc), (c, ctc), (c, gag), (c, tcg), (c, ttc), (g, acc), (g, agc), (g, atc), (g, ccc), (g, cgc), (g, gcc), (t, cgc), (t, ctc), (t, gcc)
13	(a, aac), (a, cac), (a, ccc), (a, ctc), (a, gac), (a, tac), (c, aac), (c, gac), (g, aac), (g, gac), (g, gtc), (g, tac), (t, gac)

Table 8. $\hat{P}(\cdot|I_i)$, $i = 1, \ldots, 13$, for the model – (ii) with $o = 3$ and $G = 9$, see Table 6. In bold the highest probabilities by part.

Part	a	c	g	t
1	**0.3561**	0.2003	0.2040	0.2396
2	0.2734	0.2010	0.2436	**0.2820**
3	**0.2898**	**0.2898**	0.2122	0.2083
4	**0.3599**	0.1690	0.1547	0.3165
5	0.2344	0.1285	**0.3656**	0.2715
6	**0.3019**	0.1451	0.2512	**0.3019**
7	0.2855	0.2725	0.1220	**0.3199**
8	0.2413	0.1666	0.1738	**0.4184**
9	0.3344	0.1759	0.0607	**0.4289**
10	0.2126	0.2266	0.2142	**0.3466**
11	0.2096	0.1321	0.2982	**0.3601**
12	**0.4196**	0.0883	0.1020	0.3901
13	**0.4365**	0.1867	0.0823	0.2945

Table 9. Complete genome sequences coming from https://www.ncbi.nlm.nih.gov/nuccore/Y. For each sequence Y are informed, its *Version*, sample size n, *Organism* and *Reference*.

Y	Version	n	Organism	Reference
AY304488	AY304488.1	29 731	Civet SARS CoV SZ16/2003	[14]
AY395003	AY395003.1	29 647	SARS Coronavirus ZS-C/2003	–
DQ412043	DQ412043.1	29 749	Bat SARS CoV Rm1/2004	[13]
FJ882957	FJ882957.1	29 720	SARS Coronavirus MA15	[15]
KY417144	KY417144.1	29 770	Bat SARS-like Coronavirus	[11]
MG772933	MG772933.1	29 802	Bat SARS-like Coronavirus	[12]
MG772934	MG772934.1	29 732	Bat SARS-like Coronavirus	[12]
NC_001846	NC_001846.1	31 357	Murine hepatitis virus	[16]
NC_004718	NC_004718.3	29 751	SARS-related Coronavirus	[17]
NC_019843	NC_019843.3	30 119	Middle East respiratory syndrome-related Coronavirus	[18]
NC_038294	NC_038294.1	30 111	Betacoronavirus England 1	–

Table 10. $d_{\max}$ values (see Definition 3.1) between each pair of sequences, $o = 3$, $G = 9$, $\alpha = 2$. In bold type, the values between MN908947.3 (Covid-19) and the other sequences, with * the smaller ones.

$z_{1,1}^{n_1}$	$z_{2,1}^{n_2}$	$d_{\max}$
AY304488.1	AY395003.1	0.0743
AY304488.1	DQ412043.1	0.2809
AY304488.1	FJ882957.1	0.0126
AY304488.1	KY417144.1	0.1837
AY304488.1	MG772933.1	0.3180
AY304488.1	MG772934.1	0.2584
AY304488.1	NC_001846.1	0.7409
AY304488.1	NC_004718.3	0.0767
AY304488.1	NC_019843.3	0.5340
AY304488.1	NC_038294.1	0.5340
AY395003.1	DQ412043.1	0.2809
AY395003.1	FJ882957.1	0.0501
AY395003.1	KY417144.1	0.1838
AY395003.1	MG772933.1	0.3181
AY395003.1	MG772934.1	0.2718
AY395003.1	NC_001846.1	0.7532
AY395003.1	NC_004718.3	0.0092
AY395003.1	NC_019843.3	0.5201
AY395003.1	NC_038294.1	0.5201
DQ412043.1	FJ882957.1	0.2809
DQ412043.1	KY417144.1	0.2497
DQ412043.1	MG772933.1	0.3996
DQ412043.1	MG772934.1	0.4529
DQ412043.1	NC_001846.1	0.6412
DQ412043.1	NC_004718.3	0.2808
DQ412043.1	NC_019843.3	0.4003
DQ412043.1	NC_038294.1	0.3861
FJ882957.1	KY417144.1	0.1837
FJ882957.1	MG772933.1	0.3180
FJ882957.1	MG772934.1	0.2718
FJ882957.1	NC_001846.1	0.7715
FJ882957.1	NC_004718.3	0.0529
FJ882957.1	NC_019843.3	0.4870
FJ882957.1	NC_038294.1	0.4823
KY417144.1	MG772933.1	0.3582
KY417144.1	MG772934.1	0.2327
KY417144.1	NC_001846.1	0.7259
KY417144.1	NC_004718.3	0.1837
KY417144.1	NC_019843.3	0.5010
KY417144.1	NC_038294.1	0.5010
MG772933.1	MG772934.1	0.0902
MG772933.1	NC_001846.1	0.6455
MG772933.1	NC_004718.3	0.3180
MG772933.1	NC_019843.3	0.7091
MG772933.1	NC_038294.1	0.6980
MG772934.1	NC_001846.1	0.6615
MG772934.1	NC_004718.3	0.2717
MG772934.1	NC_019843.3	0.5226
MG772934.1	NC_038294.1	0.5327
MN908947.3	AY304488.1	**0.3970**
MN908947.3	AY395003.1	**0.4236**
MN908947.3	DQ412043.1	**0.3812**
MN908947.3	FJ882957.1	**0.4134**
MN908947.3	KY417144.1	**0.4650**
MN908947.3	MG772933.1	**0.3101***
MN908947.3	MG772934.1	**0.2690***
MN908947.3	NC_001846.1	**0.8886**
MN908947.3	NC_004718.3	**0.4134**
MN908947.3	NC_019843.3	**0.5515**

(Continued on next page)

Table 10. (continued)

$z_{1,1}^{n_1}$	$z_{2,1}^{n_2}$	$d_{\max}$
MN908947.3	NC_038294.1	**0.5538**
NC_001846.1	NC_004718.3	0.7430
NC_001846.1	NC_019843.3	0.6946
NC_001846.1	NC_038294.1	0.6570
NC_004718.3	NC_019843.3	0.5200
NC_004718.3	NC_038294.1	0.5200
NC_019843.3	NC_038294.1	0.0154

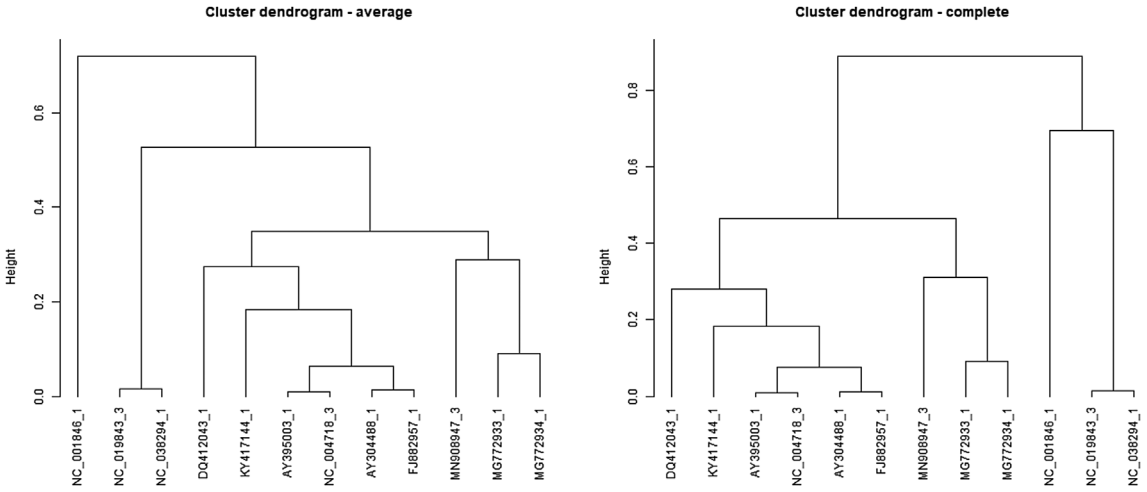

Figure 2. Dendrograms build from the $d_{\max}$ values, reported in Table 10. MN908947.3 Covid-19 sequence.

Definition 3.1. *Consider two G-Markov chains $(Z_{1,t})$ and $(Z_{2,t})$ following Definition 2.3 with alphabet A, parameters o and G, state space $\mathcal{W} = A \times A^o$ and independent samples $z_{1,1}^{n_1}$, $z_{2,1}^{n_2}$ respectively,*
(i) For $(z, s) \in \mathcal{W}$,

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

(ii)

$$d_{\max}(z_{1,1}^{n_1}, z_{2,1}^{n_2}) = \max_{(z,s)\in\mathcal{W}} \{d_{(z,s)}(z_{1,1}^{n_1}, z_{2,1}^{n_2})\},$$

with $N_{n_1+n_2}((z,s),a) = N_{n_1}((z,s),a) + N_{n_2}((z,s),a)$, $N_{n_1+n_2}(z,s) = N_{n_1}(z,s) + N_{n_2}(z,s)$, 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. With α a real and positive value.

Definition 3.1 is an adaptation of the notion introduced in [4]. It has the same properties; that is to say that i. is a metric and both i. and ii. are statistically consistent to detect if the samples come or not from the same stochastic law. Moreover, Definition 3.1-(i) is a local notion while, $d_{\max}$ – Definition 3.1-(ii) is global, so in this comparison we will use $d_{\max}$ to have a general representation of the similarity/dissimilarity between the samples, see the results in Table 10.

In Figure 2, we show the dendrograms built from $d_{\max}$ values between the pair of sequences, see Table 10. The dendrograms confirm that based on the available sequences, the sequences MN908947.3 (Covid-19), MG772934.1 and MG772933.1 could be considered as a cluster. The sequences MG772934.1 and MG772933.1 are records from July 2015 and February 2017, respectively, and the sequences come from Zhoushan China. We also see that $d_{\max}$(MG772933.1, MG772934.1) = 0.0902 is close to zero, and that proximity is confirmed in [12]. These discoveries allow us to speculate on certain aspects, one of them is that bats are consolidated as efficient transmitters and are a risk to the human immune

system, at least about Coronavirus and its last versions [13]. We note that in [10], the similarity between MN908947.3 and MG772933.1 is mentioned, and this is confirmed here, by means of the G-Model conception.

Conclusion

Partition Markov Models allow a vast economy in the construction and representation of phenomena since, through Definition 2.1, they establish units (parts) in the state space that share the same transition probabilities. Thus several states contribute to the determination of a single transition probability. The parts (elements of the process' partition) consider a finite memory o, that is to say, that the next step of the process will be determined knowing a past constituted by the concatenation of o elements coming from the alphabet. Thus, the step to time t is determined with the knowledge of the occurrences at times $t - o, \ldots, t - 1$. In this paper, we investigate a specific structure within the theoretical framework of *Partition Markov Models*. The structure of interest lies in the formulation of the partition that defines the process, in which, in addition to a finite memory o associated with the process, a parameter G is introduced, which allows dependence on the past to complement that given by the memory o, see Definition 2.3. We show how algorithms designed for the classic version of Partition Markov Models can have difficulties in recovering the structure investigated here, see Examples 2.2 and 2.3. Under previous determination of the parameters o and G it is possible to adapt all the estimation tools of the usual *Partition Markov Models* (see [1] and [4]), see Remarks 2.1 and 2.2. This specific structure in the process' partition (see Definition 2.3, Eq. (2)) is shown efficient for modeling a complete sequence of newly decoded DNA [10], Genbank MN908947, from the newly discovered Coronavirus Covid-19, from a patient of Wuhan – China. A partition in a G-model allows a huge reduction of the number of parameters to be estimated, from 768 to 39 (Tabs. 7 and 8), leading to an increase in the estimation quality of the parameters. Already, in more general terms, we see that the inclusion of the parameter G generates flexibility that is very well evaluated by model selection criteria (Tab. 6), giving credibility to partition models with more specific structures. Table 7 shows that the stochastic performance of sequence MN908947 can be reduced to 13 stochastic units that are discriminated by how the next state is selected (transition probabilities). Such a configuration could be used to design a Covid-19 profile. The model given by Definition 2.3 also allows us to develop a comparison study with 11 other genomic sequences of Coronavirus, collected in the last 25 years. We conclude that Covid-19 is shown next to Bat SARS-like Coronavirus sequences, Genbanks MG772934 and MG772933, coming from Zhoushan – China (period: 2015–2017), see Table 10 and Figure 2, see also [13]. Our results are in accordance with the indications given in [10]. This evidence could point to one of the best vectors of the virus, and help in the search for vaccines for its treatment.

Acknowledgments

G. Tasca gratefully acknowledges the partial financial support, provided by CAPES with a fellowship from the Ph.D. Program in Statistics – University of Campinas. Also, the authors wish to thank the two referees for their many helpful comments and suggestions on an earlier draft of this paper.

References

1. García Jesús E, González-López VA (2017), Consistent estimation of Partition Markov models. Entropy 19, 4, 160. https://doi.org/10.3390/e19040160.
2. Cordeiro MTA, García Jesús E, González-López VA, Mercado Londoño SL (2020), Partition Markov model for multiple processes. Math Meth Appl Sci 43, 13, 7677–7691. https://doi.org/10.1002/mma.6079.
3. Cordeiro MTA, García Jesús E, González-López VA, Mercado Londoño SL (2019), Stochastic profile of Epstein-Barr virus in nasopharyngeal carcinoma settings. 4open 2, 25.
4. García Jesús E, 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.
5. Cordeiro MTA, García Jesús E, González-López VA, Mercado Londoño SL (2019), Classification of autochthonous dengue virus type 1 strains circulating in Japan in 2014. 4open 2, 20.
6. García Jesús E, Gholizadeh R, González-López VA (2018), Stochastic distance between Burkitt lymphoma/leukemia strains, in: Demography and Health Issues, Springer, Cham, pp. 143–153.
7. García Jesús E, González-López VA (2011, November), Minimal Markov models, in: Fourth Workshop on Information Theoretic Methods in Science and Engineering, p. 25.
8. Schwarz G (1978), Estimating the dimension of a model. Ann Stat 6, 2, 461–464.
9. 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.
10. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ (2020), A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269. https://doi.org/10.1038/s41586-020-2008-3.

11. Hu B, Zeng LP, Yang XL, Ge XY, Zhang W, Li B, Xie JZ, Shen XR, Zhang YZ, Wang N, Luo DS, Zheng XS, Wang MN, Daszak P, Wang LF, Cui J, Shi ZL (2017), Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLoS Pathogens 13, 11, e1006698.
12. Hu D, Zhu C, Ai L, He T, Wang Y, Ye F, Yang L, Ding C, Zhu X, Lv R, Zhu J, Hassan B, Feng Y, Tan W, Wang C (2018), Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. Emerg Microb Infect 7, 1, 1–10.
13. Li W, Shi Z, Yu M, Ren W, Smith C, Epstein JH, Wang H, Crameri G, Hu Z, Zhang H, Zhang J, McEachern J, Field H, Daszak P, Eaton BT, Zhang S, Wang LF (2005), Bats are natural reservoirs of SARS-like coronaviruses. Science 310, 5748, 676–679.
14. Guan Y, Zheng BJ, He YQ, Liu XL, Zhuang ZX, Cheung CL, Luo SW, Li PH, Zhang LJ, Guan YJ, Butt KM, Wong KL, Chan KW, Lim W, Shortridge KF, Yuen KY, Peiris JS, Poon LL (2003), Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science 302, 5643, 276–278.
15. Roberts A, Deming D, Paddock CD, Cheng A, Yount B, Vogel L, Herman BD, Sheahan T, Heise M, Genrich GL, Zaki SR, Baric R, Subbarao K (2007), A mouse-adapted SARS-coronavirus causes disease and mortality in BALB/c mice. PLoS Pathogens 3, 1, e5.
16. Leparc-Goffart I, Hingley ST, Chua MM, Jiang X, Lavi E, Weiss SR (1997), Altered pathogenesis of a mutant of the murine coronavirus MHV-A59 is associated with a Q159L amino acid substitution in the spike protein. Virology 239, 1, 1–10.
17. He R, Dobie F, Ballantine M, Leeson A, Li Y, Bastien N, Cutts T, Andonov A, Cao J, Booth TF, Plummer FA, Tyler S, Baker L, Li X (2004), Analysis of multimerization of the SARS coronavirus nucleocapsid protein. Biochem Biophys Res Commun 316, 2, 476–483.
18. van Boheemen S, de Graaf M, Lauber C, Bestebroer TM, Raj VS, Zaki AM, Osterhaus AD, Haagmans BL, Gorbalenya AE, Snijder EJ, Fouchier RA (2012), Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. MBio 3, 6, e00473-12.

Cite this article as: García JE, González-López VA & Tasca GH 2020. Partition Markov Model for Covid-19 Virus. 4open, **3**, 13.