4open 2019, **2**, 25

https://doi.org/10.1051/fopen/2019020

Special Issue: Statistical Inference in Copula Models and Markov Processes, Case Studies and Insights
Guest Editor: Verónica Andrea González-López

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

RESEARCH ARTICLE OPEN ACCESS

Stochastic profile of Epstein-Barr virus in nasopharyngeal carcinoma settings

Marcos Tadeu Andrade Cordeiro[1], Jesús E. García[2,*], Verónica Andrea González-López[2], and Sergio Luis Mercado Londoño[2]

[1] Department of Mathematics, Federal University of Technology, Avenida Monteiro Lobato, s/n – Km 04, Ponta Grossa, CEP 84016-210 Paraná, Brazil
[2] Department of Statistics, University of Campinas, Sergio Buarque de Holanda, 651, Campinas, CEP 13083-859 São Paulo, Brazil

Received 3 March 2019, Accepted 6 June 2019

Abstract – We build a profile of the Epstein-Barr virus (EBV) by means of genomic sequences obtained from patients with nasopharyngeal carcinoma (NPC). We consider a set of sequences coming from the NCBI free source and we assume that this set is a collection of independent samples of stochastic processes related by an equivalence relation. Given a collection $\{(X_t^j)_{t\in\mathbb{Z}}\}_{j=1}^p$ of p independent discrete time Markov processes with finite alphabet A and state space S, we state that the elements (i, s) and (j, r) in $\{1, 2,\ldots, p\} \times S$ are equivalent if and only if they share the same transition probability for all the elements in the alphabet. The equivalence allows to reduce the number of parameters to be estimated in the model avoiding to delete states of S to achieve that reduction. Through the equivalence relationship, we build the global profile for all the EBV in NPC sequences, this model allows us to represent the underlying and common stochastic law of the set of sequences. The equivalence classes define an optimal partition of $\{1, 2,\ldots, p\} \times S$, and it is in relation to this partition that we define the profile of the set of genomic sequences.

Keywords: Partition Markov Models, Bayesian Information Criterion, Transition probability

Introduction

The purpose of this paper is to produce an economical model which allows representing the genomic organization of the Epstein-Barr virus (EBV), considering DNA sequences of EBV, obtained from patients with nasopharyngeal carcinoma (NPC), a disease with a distinctly high incidence in southern China. To investigate the role of EBV genomic in the pathogenesis of NPC it is necessary to describe the EBV in NPC settings. It is suggested that EBV may play a role in the development of NPC, as no other type of tumor in humans is as consistently associated with EBV as NPC. Despite the fact that EBV infection is ubiquitous, the incidence of NPC presents a remarkable geographic pattern, as it is approximately 100 times more frequent in North Africa, Southeast Asia, and Alaska than in the rest of the world. In this paper, we use two complete sequences known in the literature: GD1 [1] and GD2 [2] and an incomplete sequence, HKNPC1, reported in [3]. Phylogenetic analysis of strains in [3] also includes other sequences of EBV but not in NPC settings, and it shows that HKNPC1 is more closely related to the Chinese NPC patient-derived strains, GD1 and GD2. This evidence supports our idea to build a unique model using these three sequences.

When referring to an economical model, we are thinking about the notion introduced in [4], given in Section 2. In this article, we treat the sequences as samples of Markovian processes. By idealizing an economical model we can appeal to the principle of minimality. In the context of Markovian processes, a certain notion of minimality can be applied to the state space. Some of the ways to treat the problem are: (i) reducing the state space itself, (ii) reducing the number of probabilities to be estimated. For instance, (i) by applying deterministic finite automaton theory, to minimize an auxiliary Markov chain state space (see [5]). For (ii) by applying partition Markov models (see [6]). In the present work, we will use the second perspective, since it does not delete any state, which is relevant in our case for the characterization of EBV in NPC.

The preliminaries and the notions that we use as well as the definition of the model are given in Section 2. The model estimation is given in Section 3. We detail the database, and we show the results in Section 4, and the final considerations in Section 5.

*Corresponding author: jg@ime.unicamp.br

Preliminaries

Denote by $(X_t)_{t\in\mathbb{Z}}$ a discrete time Markov chain, with a finite alphabet and finite order. Consider $\mathcal{F} = \{(X_t^j)_{t\in\mathbb{Z}}\}_{j=1}^p$, a collection of p independent, discrete time, Markov chains on the same finite alphabet A. To simplify the notation we will assume that all the processes have the same finite memory o. $S = A^o$ is the state space of each Markov chain in the collection. Then, each string in S is a concatenation of o elements of the alphabet A. Denote the string a_m, a_{m+1},..., a_n by a_m^n, where $a_i \in A$, $m \le i \le n$. Given the process j of the collection $\mathcal{F}$ and a time t, the event $\{X_{t-o}^{jt-1} = s\}$ means that the process is equal to s in the o positions, immediately prior to the position on time t, $\{X_{t-o}^j X_{t-o+1}^j, \ldots, X_{t-1}^j = s\}$. For each $j \in J = \{1, 2,\ldots, p\}$, $a \in A$ and $s \in S$, $P^j(s) = \text{Prob}\,(X_{t-o}^{jt-1} = s)$ and $P^j(a|s) = \text{Prob}\,(X_t^j = a|X_{t-o}^{jt-1} = s)$. We define now the space where the model is established $M = J \times S$.

The parameters to be estimated in this situation are the conditional probabilities, and without the assumption of identical distribution, there are $|A|-1$ probabilities, for each state s and each process j. This corresponds to a total of $p(|A|-1)|A|^o$ parameters. Consider the situation in which there are two processes, i and j and two states s, $r \in S$ such that $P^i(a|s) = P^j(a|r)$ for all $a \in A$, then just one group of probabilities is necessary to estimate, this produces a reduction in the total number of parameters to be estimated. With this situation in mind, the following model is considered.

Definition 2.1. *The elements* (i, s), $(j, r) \in M$, *are equivalent if* $P^i(a|s) = P^j(a|r)$ *for all* $a \in A$.

This concept is very flexible in the sense that i and j could also be the same. Thus, the idea is to group all the pairs (i, s) and (j, r) that share the same probabilities and define with them, parts that constitute a partition of M.

Definition 2.2. *A collection of* p *indepedent processes* $\mathcal{F}$ *has a Markov partition* $\mathcal{L} = \{L_1, L_2,..., L_k\}$ *if* $\mathcal{L}$ *is the partition of* M *defined by the relationship introduced in Definition 2.1. Each element* L_i *of* $\mathcal{L}$ *is a part of the partition.*

Note that under the Definition 2.2, each element $(i, s) \in L$, where L is a part of $\mathcal{L}$ is such that $i \in \{1,\ldots, p\}$ and $s \in \mathcal{S}$. And note that $\mathcal{L}$ of the Definition 2.2 is minimal in the sense of having the smallest possible k, since it represents the relation of equivalence given in the Definition 2.1. Furthermore, once $\mathcal{L}$ is identified, we can also identify the conditional probabilities of the $\mathcal{F}$ collection, effectively reducing the number of parameters of the model. Given $\mathcal{L}$, we will have the following collection of parameters,

$$P(a|L_i), \quad a \in A, \quad i = 1, \ldots, k. \tag{1}$$

Then, the total number of parameters to estimate is $|\mathcal{L}|(|A| - 1)$. The process of estimating $\mathcal{L}$ requires a strategy, because as we will see, the identification of the *minimal* partition can be a fairly exhaustive process.

In the example given below we expose the notion introduced in the Definition 2.1.

Example 2.1. *Consider three processes with alphabet* $A = \{0, 1\}$ *and conditional probabilities* $P^i(\cdot|s)$, $i = 1, 2, 3$ *given by Table 1 (left) with* $s \in \mathcal{S} = A^2$. *We report the parts which compound the partition* $\mathcal{L}$ *of* $M = \{1, 2, 3\} \times \mathcal{S}$ *in the Table 1 (right).*

The number of probabilities considering the three processes separately is 12 (Table 1 – left), while by the identification made by the Definition 2.1 the number of probabilities becomes 5 (Table 1 – right).

Model estimation

Let $\left\{x_1^{Jn_j}\right\}_{j=1}^p$ be samples of the processes $\{(X_t^j)_{t\in\mathbb{Z}}\}_{j=1}^p$, with sample sizes $\{n_j\}_{j=1}^p$. Then, $\{x_1^{jn_j}\}_{j=1}^p$ can constitute a collection of independent and identically distributed realizations of one single stochastic process or only be a collection of independent realizations, coming from different processes. This means that the usual assumption (in statistical estimation) of the existence of an underlying and single law is replaced by the notion introduced in [1] and also reproduced by Definitions 2.1 and 2.2, so, the law established by Definition 2.2 is the law of the set. The number of occurrences of

Table 1. *Left*: conditional probabilities for the processes 1, 2 and 3. *Right*: parts of partition $\mathcal{L}$ of M.

s	$P^1(0\|s)$	$P^2(0\|s)$	$P^3(0\|s)$	Part (L)	$P(0\|L)$
00	0.4	0.5	0.6	$L_1 = \{(1, 10), (3, 10), (3, 11)\}$	0.1
01	0.2	0.6	0.2	$L_2 = \{(1, 01), (2, 11), (3, 01)\}$	0.2
10	0.1	0.6	0.1	$L_3 = \{(1, 00)\}$	0.4
11	0.6	0.2	0.1	$L_4 = \{(2, 00)\}$	0.5
				$L_5 = \{(1, 11), (2, 01), (2, 10), (3, 00)\}$	0.6

$s \in \mathcal{S}$ in the sample $x_1^{j^{n_j}}$ is denoted by $N((j,s))$ and the number of occurrences of s followed by a in the sample $x_1^{j^{n_j}}$ is denoted by $N((j, s), a)$. Given a partition $\mathcal{L}$ as refered in Definition 2.2, the number of occurrences of elements in L is $N(L) = \sum_{(i,s)\in L} N((i, s))$, $L \in \mathcal{L}$ and the number of occurrences of elements in L followed by $a \in A$ is, $N(L, a) = \sum_{(i,s)\in L} N((i, s), a)$. The estimator based on the Bayesian Information Criterion (BIC), associated to the samples and for the partition $\mathcal{L}$ of M is,

$$\hat{\mathcal{L}} = \operatorname{argmax}_{\mathcal{L}} \mathrm{BIC}\left(\mathcal{L}, \{x_1^{j^{n_j}}\}_{j=1}^{p}\right),$$

where,

$$\mathrm{BIC}\left(\mathcal{L}, \{x_1^{j^{n_j}}\}_{j=1}^{p}\right) = \sum_{L\in\mathcal{L}} \sum_{a\in A} N(L,a) \ln\left(\frac{N(L,a)}{N(L)}\right) - \frac{(|A|-1)}{2}|\mathcal{L}| \ln\left(\sum_{j=1}^{p} n_j\right). \tag{2}$$

Based on results derived from [6], when $\min\{n_1,\ldots, n_p\} \to \infty$, eventually almost surely, $\hat{\mathcal{L}} = \mathcal{L}$ of Definition 2.2.

Definition 3.1. *Let $\{x_1^{l^{n_1}}\}_{l=1}^{p}$ be samples of the collection $\mathcal{F} = \{(X_t^l)_{t\in\mathbb{Z}}\}_{l=1}^{p}$ of p independent Markov chains of discrete time on the same finite alphabet A with finite order o. For (i, s), $(j, r) \in M = \{1,\ldots, p\} \times A^o$,* set,

$$d((i,s),(j,r)) = \frac{2}{(|A|-1)\ln(\sum_{l=1}^{p} n_l)} \sum_{a\in A} \left\{ N((i,s),a) \ln\left(\frac{N((i,s),a)}{N((i,s))}\right) + N((j,r),a) \ln\left(\frac{N((j,r),a)}{N((j,r))}\right) \right.$$
$$\left. - N(\{(i,s),(j,r)\},a) \ln\left(\frac{N(\{(i,s),(j,r)\},a)}{N(\{(i,s),(j,r)\})}\right) \right\},$$

where $N(\{(i, s), (j, r)\}) = N((i, s)) + N((j, r))$ and $N(\{(i, s), (j, r)\}, a) = N((i, s), a) + N((j, r), a)$.

This is a *metric* on M, related to the BIC criterion in the following way, the BIC criterion indicates that (i, s), $(j, r) \in M$ should be in the same part if and only if $d((i, s), (j, r)) < 1$. d is also *statistically consistent* to decide if (i, s), $(j, r) \in M$ or not (see [6]). The metric given in Definition 3.1 goes to zero, when the laws $P^i(\cdot|s)$ and $P^j(\cdot|r)$ are identical and it takes very large values when those laws are different and when the sample sizes (n_i and n_j) grow.

Given the order o, we can define the state space $\mathcal{S}$ and also the space M, and then, we can compute all the values of d in order to determine the estimator $\hat{\mathcal{L}} = \{\hat{L}_1, \ldots, \hat{L}_k\}$ of the partition $\mathcal{L}$ (Definition 2.2). From the conception given by the Definition 2.1, the elements of a part share their probabilities, so we can also estimate the conditional probabilities $P(\cdot|L)$, by means of $P(\cdot|\hat{L}_i)$, $\cdot \in A$, $i = 1, \ldots, k$. Then, we can determine all the parameters involved in the model, proposed by the Definition 2.1.

To find the parts, the use of d can be linked to different clustering criteria: average, single linkage, agglomerative, etc. The criterion used in this work is the agglomerative, described here. Suppose that the space M is given by $M = \{m_1,\ldots, m_{|M|}\}$, define,

$$(i_*, j_*) = \operatorname{argmin}\{d(m_i, m_j), \quad i \neq j, \quad i,j \in \{1,2,\ldots,|M|\}\}, \quad \text{with } d \text{ as Definition 3.1,} \tag{3}$$

if $d(m_{i_*}, m_{j_*}) < 1$, define $M = M\backslash\{m_{i_*}\} \cup \{m_{j_*}\}] \cup m_{i^*j^*}$ with $m_{i^*j^*} = \{m_{i_*}, m_{j_*}\}$ and go back to equation (3), otherwise the procedure ends. That is, by detecting the labels i^* and j^* that indicate the globally closest elements in M, if they verify $d < 1$ then, they can be inserted in the same part, becoming an element of the new space M which is the element $m_{i^*j^*}$. The process is repeated until there are no elements such that d is less than 1.

Data and results

EBV genomes

The datasets were obtained from http://www.ncbi.nlm.nih.gov/ (National Center for Biotechnology Information [NCBI]). GD1 was isolated by infecting umbilical cord mononuclear cells by EBV from saliva of a NPC patient. GD2 and HKNPC1 were direct isolates from primary NPC biopsy specimens, see Table 2.

The coding of the sequences is exemplified below, showing the beginning of the GD1 sequence,

agaattcgtcttgctctattcacccttacttttcttcttgcccgttttctttcttagtatgaatc ...

The alphabet A has cardinal $|A| = 4$ and it is composed by the four bases: adenine (a), cytosine (c), guanine (g) and thymine (t), $A = \{a, c, g, t\}$. The set J is given by $\{1, 2, 3\}$, with 1 referring to sequence GD1, 2 referring to sequence GD2 and 3 referring to sequence HKNPC1. In stochastic processes the memory o allowed is such that $o < \log_{|A|}(n)-1$, where n is the sample size, coming from the data. In this case $n = 508\ 236$ which is around the sum of the sample sizes of the three

Table 2. EBV genomes isolated from NPC patients.

Sequence	GD1	GD2	HKNPC1
Accession number	AY961628	HQ020558	JQ009376
Patient from	Guangdong, China	Guangdong, China	Hong Kong, China
Size	174 111	166 985	167 599
Reference	[1]	[2]	[3]

sequences, then $o < 8$. In the modeling problem of genomic sequences the elements of A are given in triples, so $o = 3, 6$ are the recommended orders. Our main results are coming from the estimation with order $o = 3$, since those are more easier to expose for the reader.

Estimation

In Tables 3 and 4 we show the results with order $o = 3$. In Table 3 we see the estimate $\hat{\mathcal{L}}$ of the partition $\mathcal{L}$ which is composed by 34 estimated parts. For each estimated part $\hat{L}_i, i = 1, \ldots, 34$ we also inform its composition, listings from left to right and from top to bottom according to the order (magnitude of d) as they have been included in that part. In the last

Table 3. From top to bottom, the list of the 34 estimated parts $\hat{L}_i$ (of $\hat{\mathcal{L}}$), its elements to the right and in the last column the value d^*. 1 referring to GD1, 2 referring to GD2 and 3 referring to HKNPC1, $A = \{a, c, g, t\}$, $o = 3$, $\mathcal{S} = A^o$. In bold letter the largest and the lowest values of d^*, and on the left, the parts which have the highest probabilities (>0.4) to one element of the alphabet A (see Table 4).

i of $\hat{L}_i$	Elements $((i, s) \in J \times S)$	d^*
1	(1, *aaa*), (2, *aaa*), (3, *aaa*), (1, *caa*), (2, *caa*), (1, *tag*) (3, *tag*), (3, *caa*), (2, *tag*), (1, *taa*), (2, *taa*), (3, *taa*)	**0.98269**
2	(1, *aat*), (2, *aat*), (3, *aat*), (1, *atg*), (3, *atg*), (3, *tat*), (2, *atg*), (1, *tat*) (2, *tat*), (1, *tca*), (3, *tca*), (2, *tca*), (1, *ttg*), (3, *ttg*), (2, *ttg*)	0.92276
3	(1, *gag*), (3, *gag*), (2, *gag*), (1, *gga*), (3, *gga*), (2, *gga*)	0.92093
4	(1, *acc*), (2, *acc*), (3, *acc*), (1, *agc*), (2, *agc*), (3, *agc*), (2, *tcc*), (3, *tcc*)	0.84321
5	(1, *act*), (2, *act*), (3, *act*), (1, *gtt*), (2, *gtt*), (3, *gtt*), (1, *tcg*), (2, *tcg*), (3, *tcg*)	0.67578
6	(1, *ccg*), (1, *cgg*), (2, *ccg*), (3, *ccg*)	0.64223
7	(1, *ctt*), (2, *ctt*), (3, *ctt*), (1, *tct*), (3, *tct*), (2, *tct*)	0.62441
8	(1, *agt*), (2, *agt*), (3, *agt*), (1, *cat*), (2, *cat*), (3, *cat*)	0.56022
9	(1, *att*), (2, *att*), (3, *att*), (1, *ttt*), (3, *ttt*), (2, *ttt*)	0.49989
10	(1, *gtc*), (2, *gtc*), (3, *gtc*), (1, *tgc*), (2, *tgc*), (3, *tgc*)	0.47646
11	(1, *aac*), (2, *aac*), (3, *aac*), (1, *tac*), (2, *tac*), (3, *tac*), (1, *tcc*)	0.46913
12	(2, *cgg*), (3, *cgg*), (1, *ggt*), (2, *ggt*), (3, *ggt*), (1, *cgt*), (2, *cgt*), (3, *cgt*), (1, *ggg*)	0.38571
13	(1, *gaa*), (2, *gaa*), (3, *gaa*), (1, *gta*), (2, *gta*), (3, *gta*)	0.37340
14	(1, *gca*), (2, *gca*), (3, *gca*), (1, *gtg*), (2, *gtg*), (3, *gtg*)	0.37125
15	(1, *aag*), (2, *aag*), (3, *aag*), (1, *aga*), (3, *aga*), (2, *aga*)	0.31980
16	(1, *ctg*), (2, *ctg*), (3, *ctg*), (2, *ggg*), (3, *ggg*)	0.31369
17	(1, *ata*), (2, *ata*), (3, *ata*), (1, *tta*), (3, *tta*), (2, *tta*)	0.28809
18	(1, *agg*), (3, *agg*), (2, *agg*), (1, *tgg*), (1, *tga*), (3, *tga*), (2, *tga*), (2, *tgg*), (3, *tgg*)	0.27571
19	(1, *gcg*), (3, *gcg*), (2, *gcg*), (2, *gct*), (3, *gct*), (1, *gct*)	0.22409
20	(1, *aca*), (2, *aca*), (3, *aca*), (1, *acg*), (2, *acg*), (3, *acg*)	0.19692
21	(2, *cgc*), (3, *cgc*), (1, *ggc*), (3, *ggc*), (2, *ggc*)	0.17517
22	(1, *gcc*), (2, *gcc*), (3, *gcc*)	0.16555
23	(1, *atc*), (2, *atc*), (3, *atc*), (1, *ttc*), (2, *ttc*), (3, *ttc*)	0.09806
24	(1, *cag*), (2, *cag*), (3, *cag*)	0.05832
25	(1, *gat*), (2, *gat*), (3, *gat*)	0.05429
26	(1, *ccc*), (2, *ccc*), (3, *ccc*)	0.04653
27	(1, *cgc*), (1, *ctc*), (2, *ctc*), (3, *ctc*)	0.04151
28	(1, *cct*), (2, *cct*), (3, *cct*)	0.02430
29	(1, *tgt*), (2, *tgt*), (3, *tgt*)	0.02196
30	(1, *cca*), (3, *cca*), (2, *cca*)	0.01708
31	(1, *cga*), (2, *cga*), (3, *cga*)	0.00815
32	(1, *gac*), (2, *gac*), (3, *gac*)	0.00740
33	(1, *cac*), (3, *cac*), (2, *cac*)	0.00610
34	(1, *cta*), (2, *cta*), (3, *cta*)	**0.00328**

Table 4. Estimation of $P(\cdot|L_i)$, see equation (1). $\hat{P}(\cdot|\hat{L}_i)$, $i = 1, \ldots, 34$, $\cdot = a, c, g, t$, where the parts $(\hat{L}_i)$ are display in Table 3. In bold letter the highest probabilities by line.

i of $\hat{L}_i$	a	c	g	t
1	0.29094	0.22497	**0.29585**	0.18824
2	0.19696	0.22796	**0.32376**	0.25132
3	0.21045	0.23818	**0.39757**	0.15380
4	0.25879	**0.35082**	0.16078	0.22961
5	0.15740	0.26027	**0.32587**	0.25646
6	0.13706	0.28609	**0.40758**	0.16927
7	0.11442	0.31154	**0.31454**	0.25950
8	0.17391	**0.31700**	0.29205	0.21704
9	0.19296	0.23937	0.26847	**0.29920**
10	0.21948	**0.38880**	0.14264	0.24908
11	0.29424	**0.31157**	0.17388	0.22031
12	0.14870	0.30353	**0.36500**	0.18277
13	0.24689	0.19231	**0.38697**	0.17383
14	0.18927	0.22205	**0.38787**	0.20081
15	0.24637	0.23339	**0.33710**	0.18314
16	0.17143	0.29027	**0.35458**	0.18372
17	0.27488	0.20024	**0.28785**	0.23703
18	0.21951	0.28263	**0.30619**	0.19167
19	0.13039	0.28748	**0.37820**	0.20393
20	0.19769	0.26596	**0.31227**	0.22408
21	0.20521	**0.42192**	0.18985	0.18302
22	0.23563	**0.33767**	0.19886	0.22784
23	0.24781	**0.31734**	0.12365	0.31120
24	0.20260	0.27164	**0.38444**	0.14132
25	0.15265	0.20736	**0.42153**	0.21846
26	0.20996	**0.37591**	0.20551	0.20862
27	0.17706	**0.40208**	0.18497	0.23589
28	0.09777	**0.37674**	0.33275	0.19274
29	0.16790	**0.29442**	0.26891	0.26877
30	0.16762	0.26933	**0.39988**	0.16317
31	0.18277	0.19268	**0.45501**	0.16954
32	0.25537	**0.36960**	0.19891	0.17612
33	0.22434	**0.42235**	0.20440	0.14891
34	0.23185	**0.27230**	0.25545	0.24040

column we recorded the highest value of d, denoted by d^*, which was found by the construction of the part, applying the agglomerative criterion (see last paragraph of Section 3). In Table 4 we show the conditional probabilities of each estimated part, see equation (1).

To illustrate the estimation process of each part, we will take as an example the part 31 constituted by three elements (1, *cga*), (2, *cga*), (3, *cga*). In a first stage the elements (1, *cga*), (2, *cga*) were joined, with a $d = 0.00512$ (see Definition 3.1), this is,

$$d((1, cga), (2, cga)) = 0.00512,$$

later, this group of two elements was joined to the element (3, *cga*), with a $d = 0.00815$, see the right column in Table 3. Since (1, *cga*) and (2, *cga*) are considered identical, we can joint all the ocurrences of *cga* in the sequences GD1 and GD2. Then, for each element v of the alphabet, $v \in A$, $N((1, cga), v) + N((2, cga), v)$ records the ocurrences of *cga* followed by v and $N((1, cga)) + N((2, cga))$ records the ocurrences of *cga* in the group $\{(1, cga), (2, cga)\}$. In the second stage, the metric between the group $\{(1, cga), (2, cga)\}$ and (3, *cga*) is also computed using d from Definition 3.1,

$$d\Big(\{(1, cga), (2, cga)\}, (3, cga)\Big) = 0.00815.$$

Since both values of d are lower than 1, the three processes show the same stochastic law in relation to state *cga* but processes GD1 and GD2 are even more similar in relation to that state.

We see from the last column of Table 3, that some parts show greater homogeneity between their elements, is the case of part 34. And other parts, almost reach the limit allowed $d = 1$, exposing greater diversity, for example, part 1. This exposes the relevance of having a threshold that allows us to decide, in the light of some consistent criterion, when a discrepancy is actually detected (see [4, 6]).

Table 5. For each order $o = 3, 4, 6$ records of (a) number of estimated parts, (b) value of the BIC – see equation (2), (c) first term of the BIC (term of MLL).

Order	Number of parts	BIC value	MLL value
3	34	−682 250.0	−681 579.9
4	46	−680 495.3	−679 588.8
6	141	−650 951.1	−648 172.4

In relation to the magnitud of the conditional probabilities, we see from Table 4 that there is a tendency in all the parts to choose as the next element to visit c or g, except in the case of part 9. We also see that there are few parts that have conditional probabilities > 0.4, we identify those parts in bold letter in Table 3 (left column). For example, looking at part 31, we see a greater tendency to form the composition $cgag$ ($\hat{P}(g|cga) = 0.45501$). Now seeing the last three elements gag (members of part 3) we see that although the probability of forming $gagg$ is high, it has fallen in relation to the previous composition ($\hat{P}(g|gag) = 0.39757$).

In Table 5 we compare general aspects of three adjustments, varying the order: (i) $o = 3$, (ii) $o = 4$ and (iii) $o = 6$. We record the performance of the models for each order (each line of the Table 5).

We see that as expected the value of the BIC (and the value of maximum log-likelihood [MLL]) increases as the order of the model increases, indicating a better performance. But at the same time the number of parts quadruple from order 3 to order 6, which is why we preferred to present this study with order 3. In any case we can find the results in the following link: http://www.ime.unicamp.br/~jg/spebv/.

Conclusion

Using the model proposed in [4] we find a global representation for the three genomic sequences of Epstein-Barr virus in NPC. As we see in Example 2.1, the model foresees a reduction in the total number of parameters to be estimated and, for its estimation, we use the notion 3.1, which allows us to use different samples and different states to estimate a single probability. Taking into account the evidences of other works in relation to the similarity that these sequences show (see [3]), we take advantage of the three sequences to estimate with higher quality the parameters of the model (we use several sequences and several states for that). Also we offer a representation of the dynamic of the common process, by means of the estimated partition of the state space. We identify 34 minimum units (parts) that represent the generating process of the three sequences, and those define the partition (model). We identify the states of each part that can be considered stochastic synonyms because they produce identical transition probabilities. Note that some parts are more homogeneous than others in relation to the distance between their elements (see Table 3) and, for this reason, it is very useful to have available a threshold to use together with the metric d, this threshold is $d = 1$. In relation to the magnitude of the transition probailities of each part for the elements $\{a, c, g, t\}$, we note that those are in general < 0.4, and only six parts exceed this value in transitions for c or for g (see Table 4). With this specific case of a collection of genomic sequences of EBV in NPC, we show that although differences can be recognized between the sequences, an unique stochastic profile can be defined to the collection of sequences, when the members of the collection keep common aspects, such as being EBV in NPC.

Acknowledgments

M. Cordeiro and S. Londoño gratefully acknowledge the financial support provided by CAPES with fellowships from the PhD Program in Statistics – University of Campinas. J.E. García and V.A. González-López gratefully acknowledge the support provided by the project *Inhibitory deficit as a marker of neuroplasticity in rehabilitation* grant 2017/12943-8, São Paulo Research Foundation (FAPESP). Also, the authors wish to thank the three referees for their many helpful comments and suggestions on an earlier draft of this paper.

References

1. Zeng MS, Li DJ, Liu QL, Song LB, Li MZ, Zhang RH, Yu XJ, Wang HM, Emberg I, Zeng YX (2005), Genomic sequence analysis of Epstein-Barr Virus strain GD1 from a nasopharyngeal carcinoma patient. J Virol 79, 24, 15323–15330.
2. Liu P, Fang X, Feng Z, Guo YM, Peng RJ, Liu T, Huang Z, Feng Y, Sun X, Xiong Z, Guo X, Pang SS, Wang B, Lv X, Feng FT, Li DJ, Chen LZ, Feng QS, Huang WL, Zeng MS, Bei JX, Zhang Y, Zeng YX (2011), Direct sequencing and characterization of a clinical isolate of Epstein-Barr virus from nasopharyngeal carcinoma tissue by using next-generation sequencing technology. J Virol 85, 21, 11291–11299.
3. Kwok H, Tong AH, Lin CH, Lok S, Farrel PJ, Kwong DL, Chiang AK (2012), Genomic sequencing and comparative analysis of Epstein-Barr Virus genome isolated from primary nasopharyngeal carcinoma biopsy. PLoS One 7, 5, e36939.

4. García Jesús E, Londoño SLM (2018), Optimal model for a set of Markov processes, AIP Conference Proceedings of ICNAAM 2018, 2116.
5. Martin DEK, Aston JAD (2013), Distributions of statistics of hidden state sequences through the sum-product algorithm. Methodol Comput Appl Probab 15, 4, 897–918.
6. García JE, González-López VA (2017), Consistent estimation of partition Markov models. Entropy 19, 4, 160.

Cite this article as: Cordeiro MTA, García JE, González-López VA & Londoño SLM 2019. Stochastic profile of Epstein-Barr virus in nasopharyngeal carcinoma settings. 4open, **2**, 25.