UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística e Computação Científica

MARCOS TADEU ANDRADE CORDEIRO

Métrica para Múltiplos Processos Estocásticos

Campinas

2021

Marcos Tadeu Andrade Cordeiro

Métrica para Múltiplos Processos Estocásticos

Tese apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Estatística.

Orientadora: Verónica Andrea González López

Coorientador: Jesus Enrique Garcia

Este trabalho corresponde à versão final da Tese defendida pelo aluno Marcos Tadeu Andrade Cordeiro e orientada pela Profa. Dra. Verónica Andrea González López.

Campinas

2021

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

C811m

Cordeiro, Marcos Tadeu Andrade, 1986-
Métrica para múltiplos processos estocásticos / Marcos Tadeu Andrade Cordeiro. – Campinas, SP : [s.n.], 2021.

Orientador: Verónica Andrea González-López.
Coorientador: Jesus Enrique Garcia.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Processo estocástico. 2. Métrica (Matemática). 3. Processos de Markov. 4. Sequências estocásticas. 5. Partições (Matemática). I. González-López, Verónica Andrea, 1970-. II. Garcia, Jesus Enrique, 1966-. III. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Metric for multiple stochastic processes
Palavras-chave em inglês:
Stochastic processes
Metric (Mathematics)
Markov processes
Stochastic sequences
Partitions (Mathematics)
Área de concentração: Estatística
Titulação: Doutor em Estatística
Banca examinadora:
Verónica Andrea González-López [Orientador]
Larissa Avila Matos
Valdério Anselmo Reisen
Márcio Luis Lanfredi Viola
Mariela Fernández
Data de defesa: 24-05-2021
Programa de Pós-Graduação: Estatística

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-9689-7371
- Currículo Lattes do autor: http://lattes.cnpq.br/5552757978680842

Tese de Doutorado defendida em 24 de maio de 2021 e aprovada

pela banca examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). VERÓNICA ANDREA GONZÁLEZ LÓPEZ

Prof(a). Dr(a). LARISSA AVILA MATOS

Prof(a). Dr(a). VALDÉRIO ANSELMO REISEN

Prof(a). Dr(a). MÁRCIO LUIS LANFREDI VIOLA

Prof(a). Dr(a). MARIELA FERNÁNDEZ

A Ata da Defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós-Graduação do Instituto de Matemática, Estatística e Computação Científica.

Aos meus pais
Walmir e Anália.

Agradecimentos

Agradeço aos meus pais e meus irmãos Renan e Vinicius pelo apoio de sempre.

Agradeço à minha orientadora, Profa. Verónica, primeiramente pela acolhida, num momento muito difícil do curso, e depois pela orientação, respeito e atenção sempre dispendida. Isso se estende ao meu co-orientador também, o Prof. Jesús. Muito obrigado a vocês!

Agradeço à Unicamp, mais especificamente ao IMECC, na figura dos seus Professores e Funcionários (um destaque para o pessoal do departamento de informática) que, direta ou indiretamente, contribuíram para a realização deste trabalho e do meu curso de doutorado. Um agradecimento aos professores ministrantes das disciplinas que cursei.

Agradeço aos meus colegas de turma e de curso, Bruna, Jaime, Mario, Omar, Nathalia e Ana Roberta pelo companheirismo e amizade. Também agradeço aqui a outros amigos que não são da mesma turma, mas que, do mesmo modo, fizeram parte da minha vida durante a realização do curso, dando destaque para o Francisco, a Thalita, o Sérgio (e nossos muitos cafés nas madrugadas no predinho do IMECC, conversando sobre nossas pesquisas) e o Christian.

Agradeço também meus amigos de Ponta Grossa-PR, da UTFPR, onde leciono desde abril de 2013. Aqui um agradecimento especial àqueles mais próximos o Ednei, o David e o Bonin. Também lembro aqui do meu amigo, o Washington, que conheci também na UTFPR, mas que hoje leciona na Unicamp/FCA em Limeira. A todos meu muito obrigado.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

E, por fim, não menos importante, agradeço a UTFPR, pela liberação concedida para cursar meu doutorado e, a Fundação Araucária/Capes pela bolsa. Sem isso, não teria conseguido realizar o curso.

Resumo

Nesta tese, aplicamos a métrica d_s e noções correlatas (propostas em (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2018a) e (FERNÁNDEZ et al., 2019)), derivadas a partir do BIC. Usamos tais noções para classificar realizações de processos estocásticos Markovianos, definidos num mesmo alfabeto A e possuindo uma ordem o e espaço de estados $S = A^o$ $(o, |A|, |S| < \infty)$, doravante apenas chamados de cadeias de Markov (CM). O processo de construção e algumas propriedades de d_s (e noções correlatas) são descritos. Assim, ordenamos, em relação a sua representatividade quanto às leis de formação, sequências de DNA provenientes de dados genômicos do vírus da Dengue tipo 1 (CORDEIRO et al., 2019a) e do vírus da Zika (GARCÍA et al., 2018), sendo que tais sequências foram tratadas como realizações de CM com alfabeto $A = \{a, c, g, t\}$ e $o = 3$. Portanto, apontamos em cada caso qual sequência poderia ser usada como um "padrão-ouro" do conjunto das sequências, podendo esta ser usada em comparações com outras sequências de modo a, por exemplo, detectar novas cepas do vírus. No entanto, este procedimento tem por característica ignorar, a informação contida nas sequências não escolhidas. De modo a preencher esta lacuna, desenvolvemos uma extensão da métrica d_s (proposta em (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2018a)). Para tal é proposta uma nova noção (vide (CORDEIRO et al., 2020)) definida num conjunto $M = \{1, ..., p\} \times A^o$, onde $\{1, ..., p\}$ é um conjunto de indexadores para cada uma das p realizações independentes de CM disponíveis. Na modelagem particionamos o conjunto M usando uma relação de equivalência e, mostramos como esta métrica pode ser utilizada na obtenção do Modelo de Markov de Partições (MMP), proposto em (CORDEIRO et al., 2020). Na sequência, mostramos também algumas das propriedades teóricas da métrica, dentre elas: (i) a prova de que ela é, de fato uma métrica e, portanto, M um espaço métrico (ii) a consistência estatística da métrica na estimação do MMP (CORDEIRO et al., 2020). Adiante, trouxemos duas aplicações da métrica na obtenção de MMP, para coleções de dados genômicos dos vírus Epstein-Bar (CORDEIRO et al., 2019b) e da Zika (CORDEIRO et al., 2020). Em ambas as aplicações, um modelo único que descreve de forma parcimoniosa, a lei de formação de todas as sequências, foi obtido. Como na primeira situação, este modelo único poderia ser comparado a outras sequências não utilizadas na estimação do modelo e, assim, identificar possíveis novas cepas do vírus sob investigação.

Palavras-chave: Métrica entre processos estocásticos. Processos Markovianos. Distância entre sequências. Modelo de partições.

Abstract

In this thesis, we apply the d_s metric and related notions (proposed in (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2018a) and (FERNÁNDEZ et al., 2019)), derived from the BIC. We use these notions to classify realizations of Markovian stochastic processes, defined in the same alphabet A and having an order o and state space $\mathcal{S} = A^o$ $(o, |A|, |\mathcal{S}| < \infty)$, hereinafter referred to as Markov chains (CM). The construction process and some properties of d_s (and related notions) are described. Thus, we ordered, in relation to their representativeness regarding the formation laws, DNA sequences from genomic data of the Dengue virus type 1 (CORDEIRO et al., 2019a) and the Zika virus (GARCÍA et al., 2018), and such sequences were treated as CM realizations with alphabet $A = \{a, c, g, t\}$ and $o = 3$. Therefore, we indicate in each case which sequence could be used as a "gold standard" of the set of sequences, which can be used in comparisons with other sequences in order, for example, to detect new strains of the virus. However, this procedure has the characteristic of ignoring, the information contained in the non-chosen sequences. In order to fill this gap, we developed an extension of the d_s metric (proposed in (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2018a)). To this end, a new notion is proposed (see (CORDEIRO et al., 2020)) defined in a set $M = \{1, ..., p\} \times A^o$, where $\{1, ..., p\}$ is a set of indexers for each of the available p independent CM realizations. In the modeling we partition the set M using an equivalence relation and, we show how this metric can be used to obtain the Partition Markov Model (MMP), proposed in (CORDEIRO et al., 2020). In the sequence, we also show some of the theoretical properties of the metric, among them: (i) the proof that it is, in fact, a metric and, therefore, M a metric space (ii) the statistical consistency of the metric in the estimation of the MMP (CORDEIRO et al., 2020). Lastly, we brought two applications of the metric in obtaining MMP, for collections of genomic data of the Epstein-Bar (CORDEIRO et al., 2019b) and Zika (CORDEIRO et al., 2020) viruses. In both applications, a unique model that sparingly describes the law of formation of all sequences was obtained. As in the first situation, this unique model could be compared to other sequences not used in the estimation of the model and, thus, to identify possible new strains of the virus under investigation.

Keywords: Metric between stochastic processes. Markov Processes. Distance between Strings. Partition Markov Model.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CM	Cadeira de Markov a tempo discreto, definida num alfabeto A e com memória (ordem) o e com espaço de estados $\mathcal{S} = A^o$ ($\|A\|, o, \|S\| < \infty$).
NCP	Carcinoma nasofaringeal.
EBV	Vírus Epstein-Barr.
ZV	Vírus da Zika.
BIC	Bayesian Information Criterion.
MLV	Máximo da log-pseudo-verossimilhança.
MMP	Modelo de Markov de Partição.
DENV-1	Vírus da Dengue Tipo 1.
NCBI	National Center for Biotechnology Information - <http://www.ncbi.nlm.nih.gov/>.

Lista de símbolos

$\bar{\mathcal{P}}(M)$	É o conjunto de todos os subconjuntos de M, ou seja, é o conjunto de partes (ou conjunto potência) de M.
x_γ^η	Amostra de uma CM (X_t) contendo $\eta - \gamma + 1$ elementos.
o	ordem (memória) de uma CM.
A	alfabeto de uma CM.
$\mathcal{S}$	espaço de estados de uma CM, onde $\mathcal{S} = A^o$.
$N_n(s)$	Ver Equação (2.2). Número de vezes que o estado s, $s \in \mathcal{S}$, é observado numa amostra x_1^n de uma CM (X_t).
$N_n(s, a)$	Ver Equação (2.1). Número de vezes que o estado s, $s \in \mathcal{S}$, seguido do elemento a, $a \in A$ é observado numa amostra x_1^n de uma CM (X_t).
d_s	Ver Definição 1.
d_{max}	Ver Definição 2.
$\delta(i)$	Ver Definição 3(i).
$\beta(i)$	Ver Definição 3(ii).
M	$= \{1, ..., p\} \times A^o$
$\mathcal{L}$	$= \{L_1, L_2, ..., L_{\lvert\mathcal{L}\rvert}\}$
L_k	Corresponde a k-ésima parte do partição $\mathcal{L} = \{L_1, L_2, ..., L_{\lvert\mathcal{L}\rvert}\}$.
$d((i, s), (j, r))$	Ver Definição 8.
$N((i, s))$	Ver Equação (3.2). Número de vezes que o estado s, $s \in \mathcal{S}$, é observado na amostra $x_{i,1}^{n_i}$ de uma CM $(X_{i,t})$.
$N((i, s), a)$	Ver Equação (3.2). Número de vezes que o estado s, $s \in \mathcal{S}$, seguido do elemento a, $a \in A$, é observado na amostra $x_{i,1}^{n_i}$ de uma CM $(X_{i,t})$.

Sumário

1 Introdução

Em diversas situações, um mesmo atributo é medido em instantes distintos dando origem a uma sequência (ou a um conjunto de sequências) de valores, qualitativos ou quantitativos. Por trás destas sequências existe um mecanismo aleatório de geração. Como exemplo destas situações podemos citar: a cotação diária do real frente ao dólar, o volume pluviométrico diário (em mm) de uma região (ou até mesmo de várias regiões medidas contemporaneamente), o registro do número de ocorrências de um determinado crime (ou de alguns tipos de crimes) diariamente numa cidade, uma cadeia de DNA do vírus da Dengue (ou um grupo de cadeias vindas de pacientes diferentes infectados durante um surto), uma cadeia do vírus da Zika (ou um conjunto de cadeias de DNA vindas de pessoas infectadas. Em quaisquer das situações acima apresentadas, tais sequências podem ser vistas como amostras de processos estocásticos [1].

Com base na observação de tais sequências é necessário extrair informação que permita a descrição do fenômeno gerador. Deste modo, o conhecimento das leis de formação destas sequências, ou seja, das distribuições de probabilidade associadas a estes processos estocásticos faz-se necessário. Outra questão bastante importante é a medição do quão diferente duas (ou mais) amostras de processos estocásticos são em relação aos seus mecanismos aleatórios de geração. Esta comparação pode ser feita entre duas ou mais de duas realizações. Exemplificando, sejam dois textos convenientemente codificados [2], em que cada palavra corresponde a um elemento de um alfabeto apropriado, ou duas regiões quanto ao número de ocorrências de um determinado crime ou, duas linhas de produção são "semelhantes", pode ser feito, pela comparação das leis que regem estes processos estocásticos. Em havendo diferença entre as leis estocásticas, a pergunta natural seria onde e de que modo ela ocorre. Isso pode ser feito, por exemplo, por meio da identificação dos elementos do chamado espaço de estados (definido adiante) em que as probabilidades de transição diferem.

Considerando as sequências como amostras provenientes de processos estocásticos sob um alfabeto finito e de memória finita, em (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2018a) é introduzida uma métrica d_s que nos permite quantificar o quão próximas duas sequências estão. Essa proximidade, ou afastamento, é dada em relação às suas leis probabilísticas de formação: (i) localmente, para um mesmo estado s de ambas as amostras ou (ii) de forma global, quando considera-se o máximo de d_s. Essa métrica foi empregada com sucesso em alguns trabalhos, a saber: (i) comparação

[1] Um processo estocástico é uma sequência de variáveis aleatórias (discretas ou contínuas), definidas num mesmo espaço de probabilidades, indexadas em um conjunto T que pode ser discreto ou contínuo.

[2] A contagem do número de sílabas, letras ou até mesmo a posição da sílaba tônica, para cada palavra, são possíveis formas de codificação.

de dados genômicos de linfomas de Burkitt, de pacientes diagnosticados com leucemia (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2018b), (ii) comparação de textos escritos e codificados (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2017), (iii) análise da discrepância ou similaridade entre dois processos relacionados a produção de álcool combustível, que, a luz das suas especificações técnicas, eram indistinguíveis (GARCÍA; GONZÁLEZ-LÓPEZ; ANDRADE, 2017), dentre outros.

A métrica d_s, uma vez que mede a distância entre duas sequências, pode também ser utilizada como um classificador de um conjunto de sequências. Entende-se aqui por classificador um dispositivo que classifique (ordene, ranqueie) estas sequências entre a mais e a menos representativa, no que diz respeito as suas leis de formação. Assumindo que um conjunto de p sequências sejam amostras de processos estocásticos Markovianos, com uma mesma ordem (memória) o finita e, assumindo valores num mesmo alfabeto A, também finito, um classificador envolvendo d_s foi proposto em (FERNÁNDEZ et al., 2019). Este classificador, que está em função de d_s, é a quantidade $\delta(i)$, [3] sendo esta calculada para cada uma das p sequências $x_{i,1}^{n_i}, i = 1, ..., p$.

Este classificador, da forma como foi definido, assume valores de 0 a $+\infty$ e, quanto mais (menos) próximo $\delta(i)$ está de 0 (zero), mais (menos) representativa a observação $x_{i,1}^{n_i}$ é da lei predominante no conjunto de sequências, desde que tal lei exista. Além disso, este classificador também é robusto, no sentido de não ser afetado por pequenas variações existentes na lei predominante nas sequências. Este classificador nos permite escolher dentre um conjunto de sequências, uma (ou até mesmo mais do que uma) que seja a melhor representante deste conjunto de sequências. Por exemplo, suponhamos que dispomos de um conjunto de sequências de DNA do vírus Zika provenientes de diversas localidades e necessitamos descrever o comportamento desse vírus naquelas regiões. Uma forma de fazer isto seria através da escolha de uma sequência (dentre o conjunto delas) de DNA para ser a representante deste conjunto como um todo. A quantidade $\delta(i)$ nos permite fazer este tipo de análise. Inclusive, esta abordagem foi empregada em dois trabalhos, já publicados ou aceitos para publicação, que integram a minha contribuição para o tema. O primeiro deles foi o *Similarity Between Strains of Zika From Tropical and Subtropical Regions* (GARCÍA et al., 2018) em que utilizando sequências provenientes de 12 (doze) países latino e norte-americanos, procuramos identificar cepas que poderiam vir a ser utilizadas como cepas "padrão-ouro" no estudo do vírus da Zika. O segundo trabalho foi o *Classification of autochthonous dengue virus type 1 strains circulating in Japan in 2014* (CORDEIRO et al., 2019a) em que classificamos por representatividade um conjunto de sequências genômicas completas do vírus da dengue tipo 1, provenientes de um surto ocorrido no Japão no ano de 2014. Neste ponto, apresentamos o primeiro objetivo deste trabalho, que foi a aplicação, em problemas reais, da métrica d_s, via $\delta(i)$, como um

[3] Nesta tese, utilizaremos $\delta(i)$ ao invés de $V(x_{i,1}^{n_i})$ que foi a notação utilizada em (FERNÁNDEZ et al., 2019).

classificador de realizações de processos estocásticos. Tais processos são modelados como cadeias de Markov, com ordem finita o e alfabeto finito A.

No Capítulo 2 desta tese, abordaremos o processo de construção da métrica d_s, mostraremos algumas das suas propriedades teóricas e apresentaremos como a métrica d_s pode ser utilizada como um classificador. Na sequência, abordaremos o uso da métrica em dois problemas reais (GARCÍA et al., 2018; CORDEIRO et al., 2019a) que são parte da minha contribuição para o tema.

Até este ponto, dispomos de um conjunto de sequências e, utilizando a métrica d_s, via $\delta(i)$, classificamo-as entre a mais (menos) representativa do conjunto, em relação a lei probabilística geradora destas sequências. Desta maneira, por exemplo, poderíamos utilizar a sequência mais representativa, i.e., aquela associada ao menor valor de $\delta(i)$ para ser a representante deste conjunto. Esta abordagem, apesar de ser interessante e útil, promove o descarte das demais sequências, ou seja, inevitavelmente há uma perda de informação. Esta característica nos levou ao desenvolvimento de uma nova métrica e um novo modelo, baseados no *Bayesian Information Criterion* (BIC), que busca estabelecer um modelo global a partir do qual são identificados as coincidências e divergências entre as sequências. Esta problemática deu origem ao segundo objetivo desta tese, que é a criação de uma métrica e utilização desta para estimar o modelo de Markov de partições. Essa será a temática do Capítulo 3 desta tese.

No Capítulo 2, tínhamos uma coleção de sequências, i.e., amostras de p processos estocásticos e, os nossos objetivos foram: (i) identificação de uma lei probabilística de geração "predominante" (a chamada lei *majoritária*, Definição 4), (ii) assumindo a existência da lei majoritária, classificar (ranquear) as sequências quanto a ser gerada pela lei majoritária (qual é a melhor representante, a segunda melhor representante e assim sucessivamente.). Todavia, quando escolhermos uma (ou até mesmo duas) sequência como "padrão ouro" do grupo de sequências em questão, inevitavelmente perdemos informações, uma vez que as sequências não selecionadas serão descartadas.

Assumindo que dispomos de uma coleção de sequências de uma mesma fonte, ou seja, uma coleção de amostras $\left\{x_{j,1}^{n_j}\right\}_{j=1}^{p}$ de processos estocásticos $\{(X_{j,t})\}_{j=1}^{p}$ com uma memória o (admitida igual para todos os processos, por questões de simplicidade) e definidas num mesmo alfabeto A (o, $|A| < \infty$). Supomos também que tais sequências possuam leis distintas, possuem probabilidades de transição comuns para alguns pares de estados de algumas sequências. Baseado nisso, almejamos um modelo único onde a informação de que alguns estados em algumas sequências possuam uma mesma lei seja considerada. Para esse propósito, a métrica (uma generalização de d_s) foi construída (CORDEIRO et al., 2020), sendo esta a minha contribuição teórica para o tema. Ainda em (CORDEIRO et al., 2020), além de detalharmos a construção da métrica, obtida através do BIC de um modelo de Markov de partições, apresentamos alguns resultados teóricos

desta métrica (dentre eles, a que garante que a métrica é, de fato, uma métrica e que é estatisticamente consistente quando da detecção de discrepância ou similaridade das leis probabilísticas em comparação).

De modo a mostrar a viabilidade da métrica na estimação de um modelo de Markov de partições, realizamos duas aplicações utilizando dados reais. Ambas estão em trabalhos já publicados (CORDEIRO et al., 2019b; CORDEIRO et al., 2020) e integram minha contribuição para o tema. A primeira encontra-se em *Stochastic profile of Epstein-Barr virus in nasopharyngeal carcinoma settings* (CORDEIRO et al., 2019b). A segunda, em *Partition Markov model for multiple processes* (CORDEIRO et al., 2020), em que utilizando as mesmas sequências utilizadas em (GARCÍA et al., 2018), ajustamos um modelo de Markov particionado utilizando a métrica desenvolvida.

No capítulo 3 desta tese, trataremos do processo de construção da métrica através do modelo de Markov de partições, via BIC. Mostraremos também algumas das suas propriedades teóricas, incluindo a que mostra que é de fato uma métrica. Na sequência, uma descrição das duas aplicações feitas (CORDEIRO et al., 2019b; CORDEIRO et al., 2020). O Capítulo 4 é a conclusão.

2 A métrica d_S

Suponha que dispomos de uma coleção de sequências de uma determinada fonte, ou seja, uma coleção de amostras $\left\{x_{j,1}^{n_j}\right\}_{j=1}^{p}$ de processos estocásticos $\{(X_{j,t})_{t\in\mathbb{N}}\}_{j=1}^{p}$ de memória finita, com um espaço de estados também finito $\mathcal{S}$. Uma questão que se coloca é se essas sequências são, ou não, geradas por uma mesma lei. Para esse propósito, pode ser feito uso da métrica d_s proposta em (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2018a). Essa comparação pode ser feita *localmente*, ou seja, para um determinado estado $s \in \mathcal{S}$ ou, também, de forma *global*, através do maior valor possível desses d_s, $s \in \mathcal{S}$, ou seja, $d_{max} := \max_{s\in\mathcal{S}} d_s$. Uma aplicação do uso da métrica d_s, que faz parte da minha contribuição, é dada em (GARCÍA et al., 2018).

Tendo por base as amostras que, a luz da métrica d_s, provenham de uma mesma lei, ou seja, são geradas por uma mesma lei estocástica, poderíamos determinar um único modelo para representar a lei probabilística dessas amostras. Essa noção nos leva ao uso da métrica d_s como um *classificador*, ou seja, como um meio de classificar as sequências observadas entre a mais e a menos representativas, de um grupo pré-determinado de sequências. Assim, poder-se-á decidir quais sequências fariam parte de um mesmo grupo. Essa noção deu origem ao artigo (CORDEIRO et al., 2019a) (vide apêndice), também de minha autoria.

O presente capítulo destina-se a apresentação: do processo de construção da métrica d_s (Seção 2.1), das propriedades teóricas da métrica d_s (Seção 2.2) e o uso da métrica d_s como um classificador (Seção 2.3). Além disso, o uso da métrica em dois problemas reais também é discutido nas subseções 2.4.1 e 2.4.2. O capítulo é finalizado com uma conclusão do que foi apresentado e perspectiva de futuras pesquisas.

2.1 Construção da métrica d_s

Seja $(X_t)_{t\in\mathbb{N}}$ uma cadeia de Markov a tempo discreto de ordem o e alfabeto A (o, $|A| < \infty$) [1] ambos finitos, doravante chamadas apenas de cadeias de Markov (CM). [2] A partir desse ponto, por questões de simplicidade notacional, omitiremos o índice $t \in \mathbb{N}$ de $(X_t)_{t\in\mathbb{N}}$, ou seja, utilizaremos apenas (X_t), ficando o conjunto de índices de t, a menos dito o contrário, subentendido como sendo $t \in \mathbb{N}$. Denotaremos um *string* [3] $a_n a_{n+1} ... a_m$ por a_n^m, onde $a_i \in A$; $n \leqslant i \leqslant m$. Chamaremos aqui $\mathcal{S} = A^o$ de espaço de estados, sendo

[1] $|\star|$ é o cardinal de $\star$.

[2] Quando for mencionado, nesta tese, uma cadeia de Markov (CM) subentende-se que seja uma cadeia a tempo discreto, com memória o finita e alfabeto A, também finito.

[3] Uma *string*, nesta tese, é vista como um pedaço da cadeia de Markov.

que um estado $s \in \mathcal{S}$ é um *string* composto de o elementos. Para cada $a \in A$ e $s \in \mathcal{S}$, $P(a|s) = P(X_t = a|X_{t-o}^{t-1} = s)$.

Em uma dada amostra x_1^n, proveniente do processo estocástico (X_t):

$$N_n(s,a) := \sum_{z=1}^{n-o} \mathbb{1}_{\left\{x_z^{z+o-1} = s,\ x_{z+o} = a\right\}} \quad \text{e} \tag{2.1}$$

$$N_n(s) := \sum_{z=1}^{n-o} \mathbb{1}_{\left\{x_z^{z+o-1} = s\right\}} \quad . \tag{2.2}$$

Portanto, o número de ocorrências de s seguido de a na amostra x_1^n, é representado por $N_n(s,a)$ e, o número de ocorrências do estado s na amostra x_1^n, por $N_n(s)$. Consequentemente, o estimador de máxima verossimilhança de $P(a|s)$ é dado por $\hat{P}(a|s) = \dfrac{N_n(s,a)}{N_n(s)}$.

Exemplo 1. *Considere uma CM com ordem $o = 3$ e alfabeto $A = \{0,1\}$, onde as probabilidades de transição são dadas na Tabela 1.*

Tabela 1 – Probabilidades de transição definidas para a CM do Exemplo 1

s	$P(0\|s)$
000	0,375
001	0,850
010	0,720
011	0,400
100	0,750
101	0,175
110	0,325
111	0,675

Uma realização x_1^{300}, proveniente de uma simulação, deste processo de tamanho $n = 300$ é dada por: x_1^{300} = 001000110100001011110111011101111000100100010100001011100010001000100100010000101011101100100010000000100110111011110000100101101011010001110001110001011110111011101101110010111011101110001011011101101110000100010000010111110111011100001000100010001011101110001111110011111100001011101100111101110101. Para esta realização, as quantidades calculadas $N_{300}(s,a), a \in A$ e $\hat{P}(a|s), a \in A$ são fornecidas na Tabela 2.

Sejam $x_{1,1}^{n_1}$ e $x_{2,1}^{n_2}$ duas amostras provenientes de CM $(X_{1,t})$ e $(X_{2,t})$, respectivamente. A afirmação de que duas amostras $x_{1,1}^{n_1}$ e $x_{2,1}^{n_2}$ são geradas por uma mesma lei estocástica, ou seja, $(X_{1,t}) \overset{\text{Em lei}}{=} (X_{2,t})$ pode ser equacionada da seguinte forma: dado um $s \in \mathcal{S}$ fixo e, que as probabilidades de transição [4] são estimadas por meio das duas

[4] As probabilidades de transição, nesse caso, são representadas pelas probabilidades de ocorrência de um elemento a do alfabeto A condicionada nesse s, ou seja, $P(a|s) = P(X_t = a|X_{t-o}^{t-1} = s)$.

Tabela 2 – Valores $N_{300}(s,a), a \in A$ e $\hat{P}(a|s), a \in A$ obtidos para a realização x_1^{300} (vide página 29) do processo (X_t) definido pelas probabilidades de transição dadas na Tabela 1.

s	$N_{300}(s,0)$	$N_{300}(s,1)$	$\hat{P}(0\|s) = \frac{N_{300}(s,0)}{N_{300}(s)}$	$\hat{P}(1\|s) = \frac{N_{300}(s,1)}{N_{300}(s)}$
000	13	27	$\frac{13}{13+27} = 0{,}325$	$\frac{27}{13+27} = 0{,}675$
001	28	7	$\frac{28}{28+7} = 0{,}800$	$\frac{7}{28+7} = 0{,}200$
010	20	13	$\frac{20}{20+13} = 0{,}607$	$\frac{13}{20+13} = 0{,}393$
011	9	29	$\frac{9}{9+29} = 0{,}237$	$\frac{29}{9+29} = 0{,}763$
100	27	7	$\frac{27}{27+7} = 0{,}794$	$\frac{7}{27+7} = 0{,}206$
101	5	31	$\frac{5}{5+31} = 0{,}139$	$\frac{31}{5+31} = 0{,}861$
110	14	24	$\frac{14}{14+24} = 0{,}368$	$\frac{24}{14+24} = 0{,}632$
111	29	14	$\frac{29}{29+14} = 0{,}675$	$\frac{14}{29+14} = 0{,}325$

amostras, deveríamos estimar: (i) apenas um único conjunto de probabilidades de transição para s que seja usado simultaneamente para as duas amostras? ou (ii) dois conjuntos distintos de probabilidades de transição, uma para cada amostra?

Uma forma possível e, bem estabelecida de seleção de modelos é pelo uso do BIC (*Bayesian Information Criterion*). O BIC nada mais é que o máximo da função de log-verossimilhança do modelo, penalizada por uma função do tamanho da amostra e pelo número de parâmetros sendo estimados por esse mesmo modelo. De mais a mais, para processos markovianos o BIC é consistente[5]. Por exemplo, nos trabalhos (CSISZÁR; SHIELDS, 2000), (CSISZÁR; TALATA, 2006) e (GARCÍA; GONZÁLEZ-LÓPEZ, 2017) o BIC é usado na seleção de modelos Markovianos. Formularemos essas questões do ponto de vista do BIC.

A verossimilhança da amostra x_1^n de (X_t) é dada por:

$$P(X_1^n = x_1^n) = P(X_1 = x_1, ..., X_n = x_n) \overset{\text{CM}}{=} P(X_1^o = x_1^o) \prod_{a \in A,\, q \in \mathcal{S}} P(a|q)^{N_n(q,a)} \quad . \tag{2.3}$$

Repare que o número total de parâmetros a serem estimados em (2.3) é igual a $(|A|-1)|\mathcal{S}|$. Se $A = \{a_1, a_2, ..., a_{|A|}\}$ então $\left|\{P(a|q), q \in \mathcal{S}, a \in A \backslash \{a_{|A|}\}\}\right| = (|A|-1)|\mathcal{S}|$. [6]

[5] Entende-se por consistência, a capacidade do BIC de recuperar o modelo correto, eventualmente quase certamente, quanto o tamanho amostral é grande.

[6] Como $\sum_{a \in A} P(a|q) = 1$, então apenas $(|A|-1)$ das probabilidades $P(a|q)$ precisam ser estimadas.

O máximo da log-pseudo-verossimilhança, $MLV(\mathcal{S}, x_1^n)$, obtida em (2.3) é alcançado em:

$$MLV(\mathcal{S}, x_1^n) = \sum_{a\in A,\, q\in\mathcal{S}} N_n(q,a) \ln\left(\frac{N_n(q,a)}{N_n(q)}\right) . \tag{2.4}$$

Para uma amostra x_1^n e um modelo dado, o BIC é a diferença entre o máximo da log-pseudo-verossimilhança e uma penalização. Essa penalização está em função do número de parâmetros sendo estimados e do tamanho amostral. Portanto, para uma cadeia de Markov de ordem o e uma amostra x_1^n, temos um valor de BIC dado por: [7]

$$BIC(x_1^n) = MLV(\mathcal{S}, x_1^n) - \frac{(|A|-1)}{\alpha}|\mathcal{S}|\ln(n)\,, \tag{2.5}$$

onde $\alpha = 2$ é o BIC usual (SCHWARZ, 1978).

Consideremos duas cadeias de Markov $(X_{1,t})$ e $(X_{2,t})$ independentes, com uma mesma ordem o e definidas num mesmo alfabeto A. Além disso, dado $q \in \mathcal{S}$, sejam $\{P(a|q)\}_{a\in A}$ e $\{Q(a|q)\}_{a\in A}$ os conjuntos que representam as probabilidade de transição de $(X_{1,t})$ e $(X_{2,t})$, respectivamente. Adicionalmente, consideremos:

(i) as amostras independentes $x_{1,1}^{n_1}$ e $x_{2,1}^{n_2}$ de $(X_{1,t})$ e $(X_{2,t})$, respectivamente.

(ii) $N_{n_j}(q,a)$ como sendo o número de vezes que um dado estado $q \in \mathcal{S}$ seguido do elemento $a \in A$ ocorre na amostra $x_{j,1}^{n_j}$ do processo $(X_{j,t}), j = 1, 2$.

(iii) $N_{n_j}(q)$ como sendo o número de vezes que um dado estado $q \in \mathcal{S}$ ocorre na amostra $x_{j,1}^{n_j}$ do processo $(X_{j,t}), j = 1, 2$.

(iv) Definamos $N_{n_1+n_2}(q,a) = N_{n_1}(q,a) + N_{n_2}(q,a)$ e $N_{n_1+n_2}(q) = N_{n_1}(q) + N_{n_2}(q)$.

De início, comecemos por calcular o BIC para o modelo conjunto, sob a hipótese de que $P(.|q) \neq Q(.|q)$, $\forall q \in \mathcal{S}$. Em sendo as amostras $x_{1,1}^{n_1}$ e $x_{2,1}^{n_2}$ independentes, a verossimilhança das duas amostras é dada por:

$$\begin{aligned} P(X_{1,t}^{n_1} = x_{1,1}^{n_1}, X_{2,t}^{n_2} = x_{2,1}^{n_2}) &\overset{\text{Indep.}}{=} P(X_{1,t}^{n_1} = x_{1,1}^{n_1}) Q(X_{2,t}^{n_2} = x_{2,1}^{n_2}) \overset{\text{cadeia de Markov}}{=} \\ &= P(X_{1,t}^{o} = x_{1,1}^{o}) Q(X_{2,t}^{o} = x_{2,1}^{o}) \prod_{a\in A,\, q\in\mathcal{S}} P(a|q)^{N_{n_1}(q,a)} Q(a|q)^{N_{n_2}(q,a)}\,, \end{aligned} \tag{2.6}$$

onde, a quantidade de parâmetros sendo estimadas em (2.6) é dada por

$$n^\star = (|A|-1)|\mathcal{S}| + (|A|-1)|\mathcal{S}| = 2(|A|-1)|\mathcal{S}|\,. \tag{2.7}$$

[7] Esse critério pode ser modificado para diferentes famílias de modelos Markovianos.

Além disso, MLV$(\mathcal{S}, x_{1,1}^{n_1}, x_{2,1}^{n_2}, \neq)$, ou seja, o máximo da log-pseudo-verossimilhança para duas amostras, dada por (2.6), é igual a:

$$MLV(\mathcal{S}, x_{1,1}^{n_1}, x_{2,1}^{n_2}, \neq) = \sum_{a\in A,\, q\in\mathcal{S}} \left\{ N_{n_1}(q,a) \ln\left(\frac{N_{n_1}(q,a)}{N_{n_1}(q)}\right) + N_{n_2}(q,a) \ln\left(\frac{N_{n_2}(q,a)}{N_{n_2}(q)}\right)\right\}. \tag{2.8}$$

O BIC, $BIC(\mathcal{S}, x_{1,1}^{n_1}, x_{2,1}^{n_2}, \neq)$, associado a (2.8) é dado por:

$$BIC\left(\mathcal{S}, x_{1,1}^{n_1}, x_{2,1}^{n_2}, \neq\right) = \sum_{a\in A, q\in\mathcal{S}} \left\{ N_{n_1}(q,a) \ln\left(\frac{N_{n_1}(q,a)}{N_{n_1}(q)}\right) + \right.$$
$$\left. + N_{n_2}(q,a) \ln\left(\frac{N_{n_2}(q,a)}{N_{n_2}(q)}\right)\right\} - \frac{n^\star}{\alpha} \ln(n_1+n_2)\,. \tag{2.9}$$

onde $n^\star$ é dado por (2.7).

Por fim, calculemos o BIC para o modelo conjunto, sob a hipótese de que existe um estado $b \in \mathcal{S}$ sobre o qual $P(.|b) = Q(.|b)$. Em sendo as amostras $x_{1,1}^{n_1}$ e $x_{2,1}^{n_2}$ independentes, a verossimilhança das duas amostras é dada por:

$$\begin{aligned} P(X_{1,t}^{n_1} = x_{1,1}^{n_1}, X_{2,t}^{n_2} = x_{2,1}^{n_2}) &\overset{\text{Indep.}}{=} P(X_{1,t}^{n_1} = x_{1,1}^{n_1}) Q(X_{2,t}^{n_2} = x_{2,1}^{n_2}) \overset{\text{cadeia de Markov}}{=} \\ &= P(X_{1,t}^{o} = x_{1,1}^{o}) Q(X_{2,t}^{o} = x_{2,1}^{o}) \prod_{a\in A} P(a|b)^{N_{n_1}(b,a)} Q(a|b)^{N_{n_2}(b,a)} \\ &\quad \prod_{a\in A, q\in\mathcal{S}\setminus\{b\}} P(a|q)^{N_{n_1}(q,a)} Q(a|q)^{N_{n_2}(q,a)} = \\ &= P(X_{1,t}^{o} = x_{1,1}^{o}) Q(X_{2,t}^{o} = x_{2,1}^{o}) \prod_{a\in A} P(a|b)^{N_{n_1}(b,a)+N_{n_2}(b,a)} \\ &\quad \prod_{a\in A, q\in\mathcal{S}\setminus\{b\}} P(a|q)^{N_{n_1}(q,a)} Q(a|q)^{N_{n_2}(q,a)}\,. \end{aligned} \tag{2.10}$$

No entanto, se considerarmos a existência do estado $b \in \mathcal{S}$ sobre o qual $P(.|b) = Q(.|b)$, a quantidade de parâmetros a serem estimados, tendo por base a verossimilhança dada por (2.10), é igual a

$$n^{\star\star} = (|A|-1)|\mathcal{S}| + (|A|-1)|\mathcal{S}| - (|A|-1) = (|A|-1)(2|\mathcal{S}|-1). \tag{2.11}$$

Ademais, sob esta restrição, o máximo da log-pseudo-verossimilhança, dada por (2.10), é igual a:

$$MLV(\mathcal{S}, x_{1,1}^{n_1}, x_{2,1}^{n_2}, =_b) = \sum_{a\in A, q\in\mathcal{S}\setminus\{b\}} \left\{ N_{n_1}(q,a) \ln\left(\frac{N_{n_1}(q,a)}{N_{n_1}(q)}\right) + N_{n_2}(q,a) \ln\left(\frac{N_{n_2}(q,a)}{N_{n_2}(q)}\right)\right\} +$$
$$\sum_{a\in A} N_{n_1+n_2}(b,a) \ln\left(\frac{N_{n_1+n_2}(b,a)}{N_{n_1+n_2}(b)}\right)\quad. \tag{2.12}$$

Já o BIC associado a (2.12) é dado por:

$$BIC(\mathcal{S}, x_{1,1}^{n_1}, x_{2,1}^{n_2}, =_b) = \sum_{a\in A, q\in\mathcal{S}\setminus\{b\}} \left\{ N_{n_1}(q,a) \ln\left(\frac{N_{n_1}(q,a)}{N_{n_1}(q)}\right) + N_{n_2}(q,a) \ln\left(\frac{N_{n_2}(q,a)}{N_{n_2}(q)}\right)\right\} + \sum_{a\in A} N_{n_1+n_2}(b,a) \ln\left(\frac{N_{n_1+n_2}(b,a)}{N_{n_1+n_2}(b)}\right) - \frac{n^{\star\star}}{\alpha}\ln(n_1+n_2) \quad . \tag{2.13}$$

onde $n^{\star\star}$ é dado por (2.11).

Neste momento, dispomos de dois modelos (e seus respectivos BICs): um construído sob a hipótese de que $P(\cdot|q) \neq Q(\cdot|q)\,, \forall q \in \mathcal{S}$ e, o outro, considerando que existe um $b \in \mathcal{S}$ onde $P(\cdot|b) = Q(\cdot|b)$. Tendo em mente que o modelo com o maior BIC é o ótimo, fazendo a diferença entre os BIC de ambos os modelos, (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2018a) obtiveram uma métrica d_s. Essa quantidade, quando avaliada neste $b \in \mathcal{S}$, nos permite dizer o quão próximo ou distante as leis geradoras dos processos estão.

Definição 1. *Considere duas cadeias de Markov $(X_{1,t})$ e $(X_{2,t})$ com amostras $x_{k,1}^{n_k}$, para $k = 1, 2$ respectivamente. Para $s \in \mathcal{S}$ (com $N_{n_1}(s) > 0$ e $N_{n_2}(s) > 0$),*

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

onde $N_{n_1+n_2}(s,a) = N_{n_1}(s,a) + N_{n_2}(s,a)$, $N_{n_1+n_2}(s) = N_{n_1}(s) + N_{n_2}(s)$, N_{n_1} e N_{n_2} são dados como usualmente, obtidos a partir das amostras $x_{1,1}^{n_1}$ e $x_{2,1}^{n_2}$ respectivamente. A quantidade α é uma constante real positiva.

Exemplo 2. *Considere, novamente, uma CM com ordem $o = 3$ e alfabeto $A = \{0, 1\}$, onde as probabilidades de transição são dadas na Tabela 1 contida no Exemplo 1.*

Uma segunda realização y_1^{250}, também vinda de uma simulação, deste processo de tamanho $n = 250$ é dada por: y_1^{250} = 1100010000101101101111011101100001000010010110110110110011010001001000000001000010000101101101101110010000101000111001011011100100010000100010000010000100001000010000100001001101110110111010001011110001000100010001000010000010111011000101110111001010001000100010001000010000010111011000101110111001010. *Para esta realização, as quantidades calculadas $N_{250}(s,a), a \in A$ e $\hat{P}(a|s), a \in A$ são fornecidas na Tabela 3.*

Tabela 3 – Valores $N_{250}(s,a), a \in A$ e $\hat{P}(a|s), a \in A$ obtidos para a realização y_1^{250} (vide página 33) do processo (Y_t) definido pelas probabilidades de transição dadas na Tabela 1.

s	$N_{250}(s,0)$	$N_{250}(s,1)$	$\hat{P}(0\|s) = \frac{N_{250}(s,0)}{N_{250}(s)}$	$\hat{P}(1\|s) = \frac{N_{250}(s,1)}{N_{250}(s)}$
000	21	27	$\frac{21}{21+27} = 0{,}438$	$\frac{27}{21+27} = 0{,}562$
001	33	3	$\frac{33}{33+3} = 0{,}917$	$\frac{3}{33+3} = 0{,}083$
010	27	9	$\frac{27}{27+9} = 0{,}750$	$\frac{9}{27+9} = 0{,}250$
011	16	10	$\frac{16}{16+10} = 0{,}615$	$\frac{10}{16+10} = 0{,}385$
100	27	9	$\frac{27}{27+9} = 0{,}750$	$\frac{9}{27+9} = 0{,}250$
101	3	23	$\frac{3}{3+23} = 0{,}115$	$\frac{23}{3+23} = 0{,}885$
110	9	18	$\frac{9}{9+18} = 0{,}333$	$\frac{18}{9+18} = 0{,}667$
111	10	2	$\frac{10}{10+2} = 0{,}833$	$\frac{2}{10+2} = 0{,}167$

Assim, $d_s(x_1^{300}, y_1^{250})$ pode ser calculada para cada um dos $|\mathcal{S}| = 8$ estados $s \in \mathcal{S}$ e, seus valores são mostrados na Tabela 4.

Tabela 4 – Valores $d_s(x_1^{300}, y_1^{250})$ (ver Definição 1) obtidos entre as realizações x_1^{300} e y_1^{250} dos processos (X_t) e (Y_t), respectivamente, definido pelas probabilidades de transição dadas na Tabela 1, usando um $\alpha = 2$.

s	$d_s(x_1^{300}, y_1^{250})$
000	0,0042632
001	0,0091020
010	0,0075683
011	0,0477617
100	0,0008775
101	0,0003950
110	0,0004285
111	0,0109374

Observe que os valores de $d_s(x_1^{300}, y_1^{250})$ obtidos (Tabela 4) são todos próximos de zero. Isso era esperado uma vez que as leis geradoras de x_1^{300} e y_1^{250} são iguais (vide Tabela 1).

Repare que, para um mesmo par de amostras $x_{1,1}^{n_1}$ e $x_{2,1}^{n_2}$, obtemos $|S|$ valores de d_s. Assim, de modo a ter uma medida global de proximidade entre as leis geradoras, podemos utilizar o maior valor dentre os $|S|$ valores de d_s.

Definição 2. *Sob as mesmas suposições da Definição 1,*

$$d_{max}(x_{1,1}^{n_1}, x_{2,1}^{n_2}) = max\left\{d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}), s \in S\right\}. \tag{2.14}$$

Exemplo 3. *Consideremos ainda o Exemplo 2. O valor de d_{max} é dado por $d_{max} = 0,0477617$ que, nesse caso, é o valor de $d_{011}(x_1^{300}, y_1^{250})$. Assim, a maior distância, em lei, entre as amostras, ocorre no estado $s = \{011\}$.*

2.2 Propriedades teóricas de d_s e d_{max}

Nessa seção, apresentaremos algumas das propriedades teóricas das quantidades d_s e $d_{\max}$. Inicialmente, apresentaremos o resultado que garante d_s ser uma métrica (teorema 1). Em seguida, mostraremos a relação entre o BIC e a métrica d_s, fixado um estado $s \in \mathcal{S}$, para determinar se as probabilidades condicionais em s são iguais para os processos $(X_{1,t})$ e $(X_{2,t})$ (teorema 2). Por fim, resultados garantindo a consistência local de d_s (teorema 3) e, a consistência global de $d_{\max}$ (teorema 4) também são mostrados.

Teorema 1. *(Vide (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2018a)) Considere três amostras independentes $x_{j,1}^{n_j}$ de CM $(X_{j,t})$, $j = 1, 2, 3$. Considere um estado $s \in \mathcal{S}$ tal que $N_{n_j}(s) > 0$, para $j = 1, 2, 3$. Então:*

(i) $d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) \geqslant 0$ *com igualdade se, e somente se* $\dfrac{N_{n_1}(s,a)}{N_{n_1}(s)} = \dfrac{N_{n_2}(s,a)}{N_{n_2}(s)}$, $\forall a \in A$;

(ii) $d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) = d_s(x_{2,1}^{n_2}, x_{1,1}^{n_1})$; *e*

(iii) $d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) \leqslant d_s(x_{1,1}^{n_1}, x_{3,1}^{n_3}) + d_s(x_{3,1}^{n_3}, x_{2,1}^{n_2})$ *(desigualdade triangular).*

Prova: Vide teorema 1 de (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2018a). □

Teorema 2. *(Vide (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2018a)) Sejam duas CM $(X_{1,t})$ e $(X_{2,t})$ de ordem o. Denotemos por $\{P(a|r)\}_{a\in A, r\in\mathcal{S}}$ e $\{Q(a|r)\}_{a\in A, r\in\mathcal{S}}$ os conjuntos das probabilidades condicionais de $(X_{1,t})$ e $(X_{2,t})$ e amostras independentes $x_{1,1}^{n_1}$ e $x_{2,1}^{n_2}$, respectivamente.*

Dado um estado $s \in \mathcal{S}$ tal que $N_{n_i}(s) > 0$, para $i = 1, 2$, então:

$$BIC(x_{1,1}^{n_1}, x_{2,1}^{n_2}, \neq) < BIC(x_{1,1}^{n_1}, x_{2,1}^{n_2}, =_s) \Leftrightarrow d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) < 1$$

sendo $BIC(x_{1,1}^{n_1}, x_{2,1}^{n_2}, \neq)$ e $BIC(x_{1,1}^{n_1}, x_{2,1}^{n_2}, =_s)$ definidos em (2.9) e (2.13), respectivamente.

Prova: Vide teorema 2 de (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2018a). □

Teorema 3. *(Vide (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2018a)) (Consistência local). Em sendo as leis estocásticas de $(X_{1,t})$ e $(X_{2,t})$ as mesmas em s, então*

$$\lim_{min(n_1,n_2)\to\infty} d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) = 0 \ .$$

Caso contrário,

$$\lim_{min(n_1,n_2)\to\infty} d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) = \infty.$$

Prova: Vide teorema 3 de (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2018a). □

Lema 1. *A quantidade $d_{max} = \max\{d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}), s \in \mathcal{S}\} = 0$ se, e somente se, $d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) = 0, \forall s \in \mathcal{S}$.*

Prova: A quantidade d_s possui como menor valor o 0, ou seja, se o máximo é 0, então, todos os d_s são iguais a zero. □

Teorema 4. *(Consistência global). Em sendo as leis estocásticas de $(X_{1,t})$ e $(X_{2,t})$, iguais para todo $s \in \mathcal{S}$ então*

$$\lim_{min(n_1,n_2)\to\infty} d_{max}(x_{1,1}^{n_1}, x_{2,1}^{n_2}) = 0 \ .$$

Caso contrário,

$$\lim_{min(n_1,n_2)\to\infty} d_{max}(x_{1,1}^{n_1}, x_{2,1}^{n_2}) = \infty \ .$$

Prova: Aplicação imediata do teorema 3 e do Lema 1. □

O teorema 3 nos fornece um resultado onde podemos comparar o quão próxima ou distantes, as leis geradoras de dois processos estão em relação a um estado $s \in \mathcal{S}$. Já o teorema 4 nos permite fazer o mesmo tipo de comparação, no entanto, de forma global, ou seja, para todos os estados $s \in \mathcal{S}$. Uma aplicabilidade para o teorema 4 seria a comparação das leis estocásticas de mais do que dois processos, onde essa comparação seria feita duas a duas.

A título de exemplo, a quantidade d_{max} poderia ser calculada para pares de processos e, usando uma matriz cujas entradas fossem esses valores, poder-se-ia realizar uma análise de agrupamento (*cluster*) usando uma função de ligação conveniente (por exemplo, simples, completa, média, Ward, entre outras).

2.3 Uso da métrica d_s como um classificador

Em diversas situações práticas dispomos, não apenas de duas, mas sim de várias realizações de um processo estocástico. Por exemplo: cadeias de DNA de um vírus presente num indivíduo, valores de papeis negociados numa bolsa de valores, textos de diferentes

livros escritos por um mesmo autor, etc, podem ser vistos como realizações de processos estocásticos. Nesse contexto, duas perguntas são colocadas: (i) o quão semelhantes são as realizações de processos estocásticos em relação a um estado? (ii) assumindo um grupo de amostras, qual seria a realização mais (e a menos) representativa da lei geradora destas amostras? Estas questões serão respondidas nesta seção.

2.3.1 Seleção robusta de amostras

Tendo por base o valor de $d_{\max}$ (Definição 2), somos capazes de medir a proximidade (em termos de lei de formação) de uma determinada amostra $x_{i,1}^{n_i}$ para uma coleção de amostras $\left\{x_{j,1}^{n_j}\right\}_{\substack{j=1\\ j\neq i}}^{p}$. Adiante, com base na Definição 3, veremos uma forma de escolher uma amostra que melhor represente a lei estocástica geradora de um grupo de amostras, assumindo a existência de uma lei estocástica geradora "dominante".

Definição 3. *Dada uma coleção finita de amostras $\left\{x_{j,1}^{n_j}\right\}_{j=1}^{p}$ vindas de uma família de processos $\{(X_{j,t})\}_{j=1}^{p}$ com probabilidades $\left\{P^j(a|s), a \in A, s \in \mathcal{S}\right\}_{j=1}^{p}$, assumindo valores num alfabeto finito A, com espaço de estados $\mathcal{S} = A^o$ $(|A|, o < \infty)$. Para uma amostra $i \in \{1, 2, ..., p\}$ fixada, definimos:*

(i)

$$\delta(i) := mediana_{j\neq i}\left\{d_{max}(x_{i,1}^{n_i}, x_{j,1}^{n_j})\right\}; \tag{2.15}$$

(ii)

$$\beta(i) := mediana_{j\neq i}\left\{média_{\,s\in\mathcal{S}}\left\{d_s(x_{i,1}^{n_i}, x_{j,1}^{n_j})\right\}\right\}. \tag{2.16}$$

Observação 1. *A métrica d_s possui 0 como sendo seu valor mínimo (teorema 1 (i)). Portanto, $d_{max}(x_{i,1}^{n_i}, x_{j,1}^{n_j}) = 0$ se, e somente se $média_{\,s\in\mathcal{S}}\left\{d_s(x_{i,1}^{n_i}, x_{j,1}^{n_j})\right\} = 0$ e, $d_{max}(x_{i,1}^{n_i}, x_{j,1}^{n_j}) = \infty$ se, e somente se $média_{\,s\in\mathcal{S}}\left\{d_s(x_{i,1}^{n_i}, x_{j,1}^{n_j})\right\} = \infty$. Assim, $\delta(i) = 0$ se, e somente se $\beta(i) = 0$.*

Exemplo 4. *Consideremos 6 (seis) amostras provenientes de CM com ordem $o = 3$ e alfabeto $A = \{0, 1\}$. As probabilidades de transição são dadas nas Tabelas 5 e 6, respectivamente.*

Repare que as amostras $x_{1,1}^{n_1}$, $x_{2,1}^{n_2}$, $x_{3,1}^{n_3}$ e $x_{4,1}^{n_4}$ possuem uma mesma lei, do mesmo modo que $x_{5,1}^{n_5}$ e $x_{6,1}^{n_6}$ também. No entanto, a lei geradora dos grupos $\{x_{i,1}^{n_i}\}_{i=1}^{4}$ e $\{x_{i,1}^{n_i}\}_{i=5}^{6}$ são diferentes. As sequências são fornecidas nas Tabelas A.1, A.2, A.3, A.4, A.5 e A.6 no apêndice. As amostras $x_{1,1}^{n_1}$, $x_{2,1}^{n_2}$, $x_{3,1}^{n_3}$, $x_{4,1}^{n_4}$, $x_{5,1}^{n_5}$ e $x_{6,1}^{n_6}$, possuem tamanhos $n_1 = 400$, $n_2 = 401$, $n_3 = 402$, $n_4 = 403$, $n_5 = 404$ e $n_6 = 405$, respectivamente.

Tabela 5 – Probabilidades de transição da CM com ordem $o = 3$ e alfabeto $A = \{0, 1\}$ que geraram as amostras $x_{1,1}^{n_1}$, $x_{2,1}^{n_2}$, $x_{3,1}^{n_3}$ e $x_{4,1}^{n_4}$, apresentadas nas Tabelas A.1, A.2, A.3 e A.4, respectivamente.

s	$p(0\|s)$
000	0,70118
001	0,59719
010	0,61161
011	0,14090
100	0,46514
101	0,33474
110	0,75343
111	0,14299

Tabela 6 – Probabilidades de transição da CM com ordem $o = 3$ e alfabeto $A = \{0, 1\}$ que geraram as amostras $x_{5,1}^{n_5}$ e $x_{6,1}^{n_6}$, apresentadas nas Tabelas A.5 e A.6, respectivamente.

s	$p(0\|s)$
000	0,30228
001	0,73184
010	0,32151
011	0,82739
100	0,55165
101	0,67662
110	0,90635
111	0,41079

A Tabela 7 apresenta as quantidades $N_{n_j}(s, a)$, onde $s \in \mathcal{S}$ e $a \in \{0, 1\}$, que são necessárias para o cálculo de d_s e, consequentemente, de d_{max}. O valor da constante α usada no cálculo de d_s foi $\alpha = 2$.

Tendo por base as quantidades d_s calculadas para cada um dos $\binom{6}{2}$ pares, construímos matrizes de distâncias cujas entradas são: (i) d_{max} (Tabela 8), (ii) as médias entre os d_s (Tabela 9). Repare que a menor distância máxima d_{max} (0,01660) e a menor distância entre a média dos d_s (0,00446) ocorre entre $x_{1,1}^{n_1}$ e $x_{3,1}^{n_3}$.

Tabela 8 – Matriz de distância entre as amostras $x_{j,1}^{n_j}$, $\forall j \in 1, 2, ..., 6$ cujas entradas são d_{max} (ver 2).

	$x_{2,1}^{n_2}$	$x_{3,1}^{n_3}$	$x_{4,1}^{n_4}$	$x_{5,1}^{n_5}$	$x_{6,1}^{n_6}$
$x_{1,1}^{n_1}$	0,04089	0,01660	0,02028	0,14857	0,12958
$x_{2,1}^{n_2}$		0,02666	0,04462	0,12484	0,10716
$x_{3,1}^{n_3}$			0,02828	0,23691	0,21454
$x_{4,1}^{n_4}$				0,16945	0,14948
$x_{5,1}^{n_5}$					0,03377

Tabela 7 – Quantidades $N_{n_j}(s,0)$, $N_{n_j}(s,1)$ e $\hat{P}^j(0|s)$ para a amostra $x_{j,1}^{n_j}$, onde $j \in \{1,...,6\}$.

Amostra $x_{1,1}^{n_1}$			
s	$N_{n_1}(s,0)$	$N_{n_1}(s,1)$	$\hat{P}^1(0\|s)$
000	38	15	0,717
001	20	17	0,541
010	17	16	0,515
011	5	24	0,172
100	14	23	0,378
101	13	12	0,520
110	20	9	0,690
111	24	130	0,156

Amostra $x_{2,1}^{n_2}$			
s	$N_{n_2}(s,0)$	$N_{n_2}(s,1)$	$\hat{P}^2(0\|s)$
000	31	17	0,646
001	24	12	0,667
010	16	12	0,571
011	7	24	0,226
100	17	20	0,459
101	4	19	0,174
110	21	10	0,677
111	24	140	0,146

Amostra $x_{3,1}^{n_3}$			
s	$N_{n_3}(s,0)$	$N_{n_3}(s,1)$	$\hat{P}^3(0\|s)$
000	44	10	0,815
001	20	14	0,588
010	16	11	0,593
011	1	28	0,034
100	11	24	0,314
101	7	14	0,333
110	19	10	0,655
111	28	142	0,165

Amostra $x_{4,1}^{n_4}$			
s	$N_{n_4}(s,0)$	$N_{n_4}(s,1)$	$\hat{P}^4(0\|s)$
000	25	13	0,658
001	10	25	0,286
010	10	6	0,625
011	5	33	0,132
100	13	21	0,382
101	7	13	0,350
110	24	14	0,632
111	33	148	0,182

Amostra $x_{5,1}^{n_5}$			
s	$N_{n_5}(s,0)$	$N_{n_5}(s,1)$	$\hat{P}^5(0\|s)$
000	9	26	0,257
001	45	17	0,726
010	30	55	0,353
011	33	6	0,846
100	27	36	0,429
101	39	22	0,639
110	33	6	0,846
111	6	11	0,353

Amostra $x_{6,1}^{n_6}$			
s	$N_{n_6}(s,0)$	$N_{n_6}(s,1)$	$\hat{P}^6(0\|s)$
000	14	35	0,286
001	43	18	0,705
010	24	62	0,279
011	29	7	0,806
100	34	26	0,567
101	44	18	0,710
110	36	0	1,000
111	7	5	0,583

Tabela 9 – Matriz de distância entre as amostras $x_{j,1}^{n_j}$, $\forall j \in 1,2,...,6$ cujas entradas são iguais a média $_{s\in\mathcal{S}}\left\{d_s(x_{i,1}^{n_i}, x_{j,1}^{n_j})\right\}$.

	$x_{2,1}^{n_2}$	$x_{3,1}^{n_3}$	$x_{4,1}^{n_4}$	$x_{5,1}^{n_5}$	$x_{6,1}^{n_6}$
$x_{1,1}^{n_1}$	0,00651	0,00446	0,00456	0,03309	0,04622
$x_{2,1}^{n_2}$		0,00712	0,00819	0,03593	0,05132
$x_{3,1}^{n_3}$			0,00625	0,05267	0,06779
$x_{4,1}^{n_4}$				0,04464	0,05933
$x_{5,1}^{n_5}$					0,00756

Além disso, a mediana das linhas (ou colunas) das Tabelas 8 e 9, resultam nas quantidades $\delta(i)$ e $\beta(i)$, respectivamente. Por exemplo, $\delta(4)$ é a mediana das quantidades $d_{max}(x_{4,1}^{n_4}, x_{1,1}^{n_1})$, $d_{max}(x_{4,1}^{n_4}, x_{2,1}^{n_2})$, $d_{max}(x_{4,1}^{n_4}, x_{3,1}^{n_3})$, $d_{max}(x_{4,1}^{n_4}, x_{5,1}^{n_5})$ e $d_{max}(x_{4,1}^{n_4}, x_{6,1}^{n_6})$ que é igual a 0,04462, i.e., $\delta(4) = 0,04462$. Já $\beta(4)$ é a mediana das quantidades média $_{s\in\mathcal{S}}\left\{d_s(x_{4,1}^{n_4}, x_{1,1}^{n_1})\right\}$, média $_{s\in\mathcal{S}}\left\{d_s(x_{4,1}^{n_4}, x_{2,1}^{n_2})\right\}$, média $_{s\in\mathcal{S}}\left\{d_s(x_{4,1}^{n_4}, x_{3,1}^{n_3})\right\}$, média $_{s\in\mathcal{S}}\left\{d_s(x_{4,1}^{n_4}, x_{5,1}^{n_5})\right\}$ e média $_{s\in\mathcal{S}}\left\{d_s(x_{4,1}^{n_4}, x_{6,1}^{n_6})\right\}$ que é igual a $\beta(4) = 0,00819$.

As quantidades $\delta(i)$ e $\beta(i)$ são apresentadas na Tabela 10. Repare que existe um $\delta(i)$ e $\beta(i)$ associado a cada uma das amostras $x_{i,1}^{n_i}$, onde $i \in \{1, ..., 6\}$, ou seja, $\delta(4)$ e $\beta(4)$ estão associados a $x_{4,1}^{n_4}$, $\delta(2)$ e $\beta(2)$ estão associados a $x_{2,1}^{n_2}$, e assim por diante.

Após a ordenação das amostras, em ordem crescente de acordo com $\delta(i)$ e $\beta(i)$, obtivemos uma ordenação muito parecida: (i) $x_{3,1}^{n_3}$, $x_{1,1}^{n_1}$, $x_{2,1}^{n_2}$, $x_{4,1}^{n_4}$, $x_{6,1}^{n_6}$ e $x_{5,1}^{n_5}$, Tabela 10 (a) e, (ii) $x_{1,1}^{n_1}$, $x_{3,1}^{n_3}$, $x_{2,1}^{n_2}$, $x_{4,1}^{n_4}$, $x_{5,1}^{n_5}$ e $x_{6,1}^{n_6}$ ao usarmos $\beta(i)$, Tabela 10 (b). Também podemos observar pelas Tabelas 10 (a) e (b) que existe um salto, quanto ao valor de $\delta(i)$ entre as observações $x_{4,1}^{n_4}$ e $x_{6,1}^{n_6}$ na Tabela 10 (a) e, um outro salto, mas agora entre $x_{4,1}^{n_4}$ e $x_{5,1}^{n_5}$ na Tabela 10(b).

Tabela 10 – Valores de $\delta(i)$ (2.15) e $\beta(i)$ (2.16) em (a) e (b), respectivamente, ordenados crescentemente com suas respectivas amostras.

(a)

Amostra	$\delta(i)$
$x_{3,1}^{n_3}$	0,02828
$x_{1,1}^{n_1}$	0,04089
$x_{2,1}^{n_2}$	0,04462
$x_{4,1}^{n_4}$	0,04462
$x_{6,1}^{n_6}$	0,12958
$x_{5,1}^{n_5}$	0,14857

(b)

Amostra	$\beta(i)$
$x_{1,1}^{n_1}$	0,00651
$x_{3,1}^{n_3}$	0,00712
$x_{2,1}^{n_2}$	0,00819
$x_{4,1}^{n_4}$	0,00819
$x_{5,1}^{n_5}$	0,03593
$x_{6,1}^{n_6}$	0,05132

Além disso, construímos dendrogramas usando como matriz de distâncias, as matrizes fornecidas nas Tabelas 8 e 9; e usando a ligação de Ward (ver Figura 1). Em ambos os dendrogramas observamos dois grupos, a saber: (i) amostras 5 ($x_{5,1}^{n_5}$) e 6 ($x_{6,1}^{n_6}$), (ii) amostras 2 ($x_{2,1}^{n_2}$), 4 ($x_{4,1}^{n_4}$), 3 ($x_{3,1}^{n_3}$) e 1 ($x_{1,1}^{n_1}$). Os grupos estão separados por lei de acordo com as fontes geradoras.

Em situações práticas, nem sempre uma coleção finita de amostras provém de leis iguais. Mas, mesmo em tais situações, se

$$J_i = \left\{ j : 1 \leqslant j \leqslant p, P^j(a|s) = P^i(a|s), a \in A, s \in \mathcal{S} \right\},$$

então

$$\delta(i) \underset{min\{n_1,...,n_p\}\to\infty}{\longrightarrow} 0,\ i = 1, 2, ..., p \Leftrightarrow \xi_i = |J_i| > \left\lceil \frac{p}{2} \right\rceil$$

onde $\lceil \dagger \rceil$ é o menor número inteiro maior que $\dagger$. Em palavras, J_i é um conjunto que indica quais das p amostras $\left\{x_{1,1}^{n_1}, ..., x_{p,1}^{n_p}\right\} \setminus \left\{x_{i,1}^{n_i}\right\}$ possuem uma lei de formação igual a de $x_{i,1}^{n_i}$.

Na sequência, apresentamos um resultado que expõe formalmente as condições nas quais $\delta(i)$, ao assumir valores altos, indica uma proporção alta de amostras provenientes de leis estocásticas distintas.

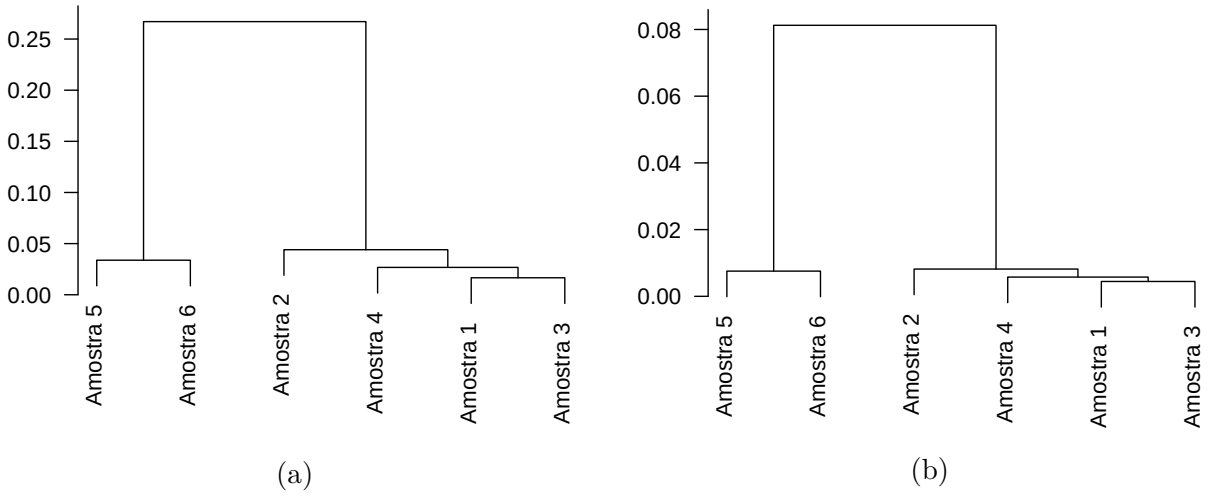


Figura 1 – Dendrograma das sequências usando o método aglomerativo hierárquico de *Ward*. As matrizes de distância entre as amostras 1 ($x^{n_1}_{1,1}$, A.1), 2 ($x^{n_2}_{2,1}$, A.2), 3 ($x^{n_3}_{3,1}$, A.3), 4 ($x^{n_4}_{4,1}$, A.4), 5 ($x^{n_5}_{5,1}$, A.5) e 6 ($x^{n_6}_{6,1}$, A.6), têm suas entradas dadas por: (a) d_{max} e (b) média $_{s\in\mathcal{S}}\left\{d_s(x^{n_i}_{i,1}, x^{n_j}_{j,1})\right\}$.

Teorema 5. *(Vide (FERNÁNDEZ et al., 2019)) Sob as condições da definição 3 e*

$$^{8}\,\xi_i = \left|\left\{j : 1 \leqslant j \leqslant p, P^j(a|s) = P^i(a|s), a \in A, s \in \mathcal{S}\right\}\right|$$

para cada $i \in \{1, ..., p\}$, $\delta(i) \underset{min\{n_1,...,n_p\}\to\infty}{\longrightarrow} \infty \Leftrightarrow \xi_i \leqslant \left\lceil \frac{p}{2} \right\rceil$.

Prova: Vide teorema 3.1 de (FERNÁNDEZ et al., 2019). □

Assim, a luz do teorema 5, $\delta(i)$ assume valores grandes se, e somente se, o número de amostras que compartilham a lei com $x^{n_i}_{i,1}$ não atinge 50%. Para que possamos extrair, de uma coleção de amostras, uma amostra específica que represente a lei do conjunto, assumiremos a existência de uma lei estocástica predominante.

Definição 4. *Sob as suposições da Definição 3 e*

$$\xi_i = \left|\left\{j : 1 \leqslant j \leqslant p, P^j(a|s) = P^i(a|s), a \in A, s \in \mathcal{S}\right\}\right|,$$

a lei $\left\{P^k(a|s), a \in A, s \in \mathcal{S}\right\}$*, para algum* $k \in \{1, ..., p\}$ *é chamada de lei majoritária em* $\{(X_{i,t})\}_{i=1}^{p}$ *se,* $\xi_k > \xi_i$, $\forall i \in \{1, ..., p\}$, $i \neq k$.

A Definição 4 chama a lei estocástica que explica a maior quantidade de processos $\{(X_{j,t})\}_{j=1}^{p}$ de lei majoritária.

[8] A quantidade ξ_i é o número de amostras, dentre as p disponíveis, que possuem a mesma lei de formação da i-ésima amostra.

Observação 2. *Se $\xi_k \leqslant \left\lceil \frac{p}{2} \right\rceil$ então, para cada $i \in \{1, ..., p\}$, $\xi_i \leqslant \left\lceil \frac{p}{2} \right\rceil$ e, de acordo com o teorema 4, $\delta(i) \to \infty$ sempre que $min\{n_1, ..., n_p\} \to \infty$.*

Assim sendo, de acordo com a Observação 2, se a lei majoritária não corresponder a, pelo menos, metade das amostras, então qualquer procedimento baseado em $\delta(i)$ será inconclusivo.

Considere uma situação em que dispomos de p amostras e, queremos identificar qual delas é, em termos de lei estocástica, a melhor representante delas. Adiante é apresentado, um procedimento para a ordenação $\mathcal{P}(\{x_{j,1}^{n_i}\}_{j=1}^{p})$ das p amostras que faz uso das quantidades δ.

Procedimento 1. $(\mathcal{P}(\{x_{j,1}^{n_i}\}_{j=1}^{p}))$

1. *Entrada (conjunto de amostras):* $\{x_{j,1}^{n_j}\}_{j=1}^{p}$
   a) *Para cada j, calcule $\delta(j)$, onde $j = 1, 2, ..., p$*
   b) *Calcule Δ_i, onde Δ_i é a i-ésima estatística de ordem de $\{\delta(j), j = 1, ..., p\}$, para $i = 1, 2, ..., p$.*
   c) *Denotemos por $x_{(i),1}^{n_{(i)}}$ a amostra relacionada a Δ_i, para $i = 1, ..., p$.*
2. *Saída (conjunto de amostras ordenadas):* $\left\{x_{(j),1}^{n_{(j)}}\right\}_{j=1}^{p}$.

Assim sendo, o Procedimento 1 transforma um conjunto de p amostras, em um novo conjunto ordenado, através do uso da quantidade δ.

Teorema 6. *(Vide (FERNÁNDEZ et al., 2019)) Considere as amostras $\{x_{j,1}^{n_j}\}_{j=1}^{p}$ vindas dos processos $\{(X_{j,t}^{n_j})\}_{j=1}^{p}$ sob as suposições da Definição 4. Seja $n_{min} = \min\{n_1, ..., n_p\}$ e ξ_k a quantidade de amostras que são geradas pela lei majoritária. Suponha que $\xi_k > \left\lceil \frac{p}{2} \right\rceil$ e defina $x_1^{n_x} = \mathcal{P}(\{x_{j,1}^{n_j}\}_{j=1}^{p})(i)$ onde $\mathcal{P}$ é o Procedimento de ordenação 1. Assim, para um n_{min} suficientemente grande e $i \leqslant \xi_k$, $x_1^{n_x}$ será uma amostra da lei majoritária.*

Prova: Vide demonstração do teorema 2 de (FERNÁNDEZ et al., 2019). □

O teorema 6 expresso em palavras. Assuma que a lei majoritária exista e que possua ξ_k casos. Após a aplicação do procedimento 1, qualquer amostra associada a i-ésima ordem será uma amostra da lei majoritária, desde que i seja no máximo igual a ξ_k.

2.4 Estudos de caso

2.4.1 Similaridade entre cepas do vírus Zika de regiões tropicais e subtropicais

Nesse ponto, apresento uma descrição do estudo (GARCÍA et al., 2018), parte integrante da minha contribuição, já publicado. Nesse trabalho, analisamos dados genômicos provenientes de cepas [9] do vírus Zika, obtidos da fonte gratuita NCBI [10]. O principal objetivo do trabalho foi identificar uma cepa que melhor representasse o conjunto de cepas, ou seja, uma espécie de "cepa padrão" que fosse capaz de representar todo o conjunto de sequências. Tais cepas podem ser usadas como perfis de cepas do vírus Zika.

Cada uma dessas cepas foi modelada por uma cadeia de Markov ergódica, de ordem finita e sob um alfabeto $A = \{a, c, g, t\}$ e a proximidade entre as cepas, aos pares, foi determinada usando-se a noção $d_{max} = \max\{d_s, s \in \mathcal{S}\}$ (Definição 2). Para a identificação de quais cepas estão próximas ou distantes entre si, tendo em vista suas leis probabilísticas de formação, usamos as quantidades $\delta(i)$ e $\beta(i)$ da Definição 3. Essas quantidades nos permitem classificar as cepas em ordens de representatividade, ou seja, qual delas é a mais representativa, a segunda mais presentativa e assim por diante, até chegar a menos representativa. Essa metodologia possibilita-nos estabelecer quais grupos mostram um padrão similar. Adicionalmente, quando fazemos referência a um comportamento padrão, estamos, obviamente, assumindo que essas sequências têm suas peculiaridades e, portanto, deveríamos fazer uso de um procedimento robusto que minimize grandes distorções. Esse último motivo nos levou ao uso de $\delta(i)$.

2.4.1.1 Dados genômicos:

Com relação aos dados, cada sequência de DNA foi considerada como sendo uma amostra de uma família de CM $\{(X_{j,t})\}_{j=1}^{p}$, ou seja, x_1^n é formada pela concatenação de elementos do alfabeto $A = \{a, c, g, t\}$. As amostras, nesse estudo, totalizando $p = 153$, provém de doze países, a saber: (1) Brasil [BRA] (44 amostras), (2) Estados Unidos da América [USA] (34 amostras), (3) Republica Dominicana [DOM] (23 amostras), (4) México [MEX] (19 mostras), (5) Honduras [HND] (13 amostras), (6) Nicarágua [NIC] (7 amostras), (7) Jamaica [JAM] (4 amostras), (8) Colômbia [COL] (3 amostras), (9) Porto Rico [PRI](2 amostras), (10) Venezuela [VEN](2 amostras), (11) Cuba [CUB] (1 amostra) e (12) Martinica [MTQ] (1 amostra). A listagem completa das sequências utilizadas neste estudo é apresentada na Tabela 18.

Nesse trabalho, além de considerarmos que todas sequências genômicas comportam-

[9] Cepa é uma variante genética ou subtipo de um micro-organismo (por exemplo, um fungo, uma bactéria ou um vírus). Ela faz referência a um grupo de descendentes que possuem um ancestral comum. Assim, acabam compartilhando semelhanças morfológicas e fisiológicas. A título de exemplo, os vírus H1N1 (inicialmente chamado de *gripe suína*), H5N1 (gripe aviária) são cepas do vírus influenza (gripe).

[10] <https://www.ncbi.nlm.nih.gov>

se como amostras de processos estocásticos, assumimos que no mínimo 50% delas seguem a mesma lei estocástica. Portanto, com essa suposição, podemos classificar as amostras entre mais representativa e menos representativa, ou seja, podemos usar o máximo da métrica d_s, ou seja, d_{max} como um classificador (FERNÁNDEZ et al., 2019) . O menor tamanho de amostra dessas sequências, ou seja, $n_{min} = min\{n_1, ..., n_{153}\}$ foi de 10807. De acordo com a regra usual que limita a memória máxima da cadeia em $o < \lfloor log_{|\mathcal{A}|}(n) \rfloor - 1 = 5$, teríamos $o = 5$. Mas, como as bases desse alfabeto são organizadas em triplas, nesse estudo consideramos $o = 3$. Além disso, o valor de α necessário para o cálculo de d_s e, consequentemente de d_{max} e $\delta(i)$ e $\beta(i)$ foi assumido como sendo igual a $\alpha = 2$ (SCHWARZ, 1978).

2.4.1.2 Resultados:

A inspeção dos resultados foi separada em três subseções: (i) as amostras referentes ao Brasil, (ii) as amostras norte-americanas, (iii) todas as amostras consideradas conjuntamente. Os dados do Brasil e EUA foram analisados, inicialmente, separados pois constituem os maiores grupos da base de dados.

Em cada uma das subseções, consideramos um universo $\mathcal{C}$ de amostras sendo comparadas. Além disso, para cada conjunto de amostras $\mathcal{C}$, as distâncias máximas d_{max} foram calculadas entre pares i e j, onde $i \neq j$ de amostras em $\mathcal{C}$. Com essas quantidades calculadas, podemos construir grupos que nos permitirão identificar quais amostras são mais próximas ou distantes, em termos de leis probabilísticas. Por fim, pelo uso da quantidade $\delta(i)$, poderemos classificar cada uma das amostras no universo $\mathcal{C}$, em mais representativa e menos representativa. Essa classificação é feita, resumidamente, do seguinte modo: (i) para cada amostra $i \in \mathcal{C}$, um valor de $\delta(i)$, calculado com base na Definição 3, é associado. (ii) O valor de $\delta(i)$ mais baixo (alto) indica a amostra mais (menos) representativa (Procedimento 1).

Sequências brasileiras A Tabela 11, adaptada de (GARCÍA et al., 2018), fornece as sequências brasileiras do vírus da Zika, seus respectivos valores calculados de $\delta(i)$ e $\beta(i)$. As ordens das estatísticas de ordem de $\delta(i)$ e $\beta(i)$ também são apresentadas. As sequências são ordenadas, de acordo com a magnitude, em ordem crescente, de $\delta(i)$.

Na Tabela 15 são apresentadas algumas estatísticas descritivas, calculadas a partir dos valores obtidos de $\delta(i)$ e $\beta(i)$ (segunda e quarta colunas). Por exemplo, os valores médios de $\delta(i)$ e $\beta(i)$ obtidos para as sequências brasileiras foram 0,04248 e 0,00930, respectivamente. Já os coeficientes de variação de $\delta(i)$ e $\beta(i)$ foram 0,86381 e 0,83750, ou seja, a distribuição dos valores de $\delta(i)$ e $\beta(i)$ estão centradas em valores diferentes, mas, possuem uma variabilidade relativa semelhante.

Tabela 11 – Da esquerda para a direita (a) sequências brasileiras de vírus Zika, (b) valor de $\delta(i)$, (c) posto de $\delta(i)$ (Definição 3 (i)) (d) valor de $\beta(i)$ (Definição 3 (ii)) e (e) posto de $\beta(i)$. Sequências ordenadas em ordem crescente, de acordo com o valor de $\delta(i)$.

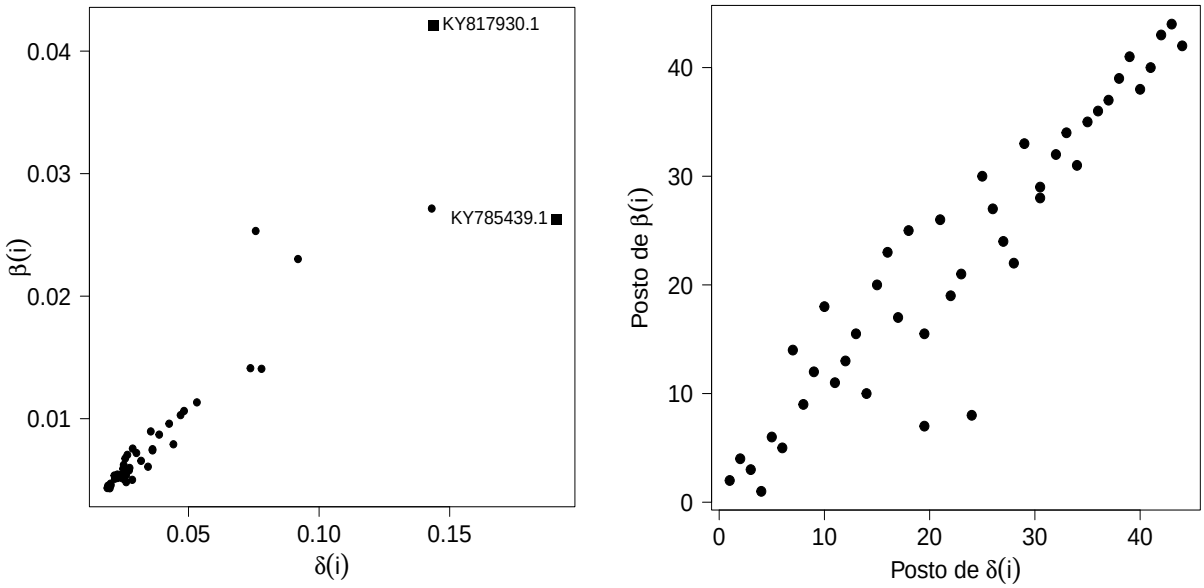
Sequência	δ	$\delta_{(k)}$	β	$\beta_{(k)}$
KY558999.1	0,01894	1	0,00434	2
KY559005.1	0,01919	2	0,00453	4
KY785450.1	0,01947	3	0,00444	3
KY559015.1	0,01987	4	0,00432	1
KY559007.1	0,02017	5	0,00470	6
KY559027.1	0,02025	6	0,00456	5
KY014307.2	0,02160	7	0,00536	14
KY785410.1	0,02180	8	0,00510	9
KY014320.2	0,02254	9	0,00523	12
KY014296.2	0,02277	10	0,00544	18
KY559013.1	0,02287	11	0,00516	11
KY785433.1	0,02335	12	0,00527	13
KY785479.1	0,02395	13	0,00539	15,5
KY785426.1	0,02471	14	0,00513	10
KY559012.1	0,02495	15	0,00591	20
KY559021.1	0,02519	16	0,00623	23
KY785455.1	0,02529	17	0,00540	17
KY785456.1	0,02576	18	0,00675	25
KY014317.2	0,02615	19,5	0,00539	15,5
KY785427.1	0,02615	19,5	0,00482	7
KY559023.1	0,02660	21	0,00705	26
KY014301.2	0,02725	22	0,00578	19
KX197192.1	0,02745	23	0,00598	21
KY014297.2	0,02847	24	0,00501	8
KY559024.1	0,02863	25	0,00756	30
KY559017.1	0,03005	26	0,00720	27
KY559003.1	0,03185	27	0,00656	24
KY785429.1	0,03450	28	0,00608	22
KY559006.1	0,03556	29	0,00896	33
KY559019.1	0,03623	30,5	0,00741	28
KY559018.1	0,03623	30,5	0,00751	29
KY014313.2	0,03878	32	0,00870	32
KY559031.1	0,04260	33	0,00959	34
KY559011.1	0,04423	34	0,00791	31
KY785437.1	0,04697	35	0,01029	35
KY559032.1	0,04830	36	0,01063	36
KY559010.1	0,05322	37	0,01133	37
KY559009.1	0,07372	38	0,01412	39
KY014308.2	0,07573	39	0,02532	41
KY559001.1	0,07797	40	0,01407	38
KY559014.1	0,09192	41	0,02303	40
KY559004.1	0,14313	42	0,02715	43
KY817930.1	0,14382	43	0,04206	44
KY785439.1	0,19106	44	0,02623	42

Fonte: (GARCÍA et al., 2018).

Ainda na Tabela 11, os postos de $\delta(i)$ e $\beta(i)$ também são apresentados, nas colunas $\delta_{(k)}$ e $\beta_{(k)}$, respectivamente. Por exemplo, o valor de $\delta(i)$ associado a amostra KY558999.1 ($\delta = 0,01894$) corresponde ao $\delta(i)_{(1)}$, ou seja, o mínimo. Já o $\beta(i)$ correspondente a esta mesma amostra ($\beta = 0,00434$) é o $\beta(i)_{(2)}$, ou seja, o segundo menor valor de $\beta(i)$.

Repare que existe certa concordância entre: as quantidades $\delta(i)$ e $\beta(i)$ (Figura 2 (a)) e, os postos de $\delta(i)$ e $\beta(i)$ (Figura 2 (b)). A correlação amostral entre $\delta(i)$ e $\beta(i)$ obtida foi de 0,9099. Já a correlação amostral entre os postos de $\delta(i)$ e $\beta(i)$ foi igual a 0,9385. Portanto, existe uma associação linear positiva entre δ e β e, também, entre os postos destas quantidades.

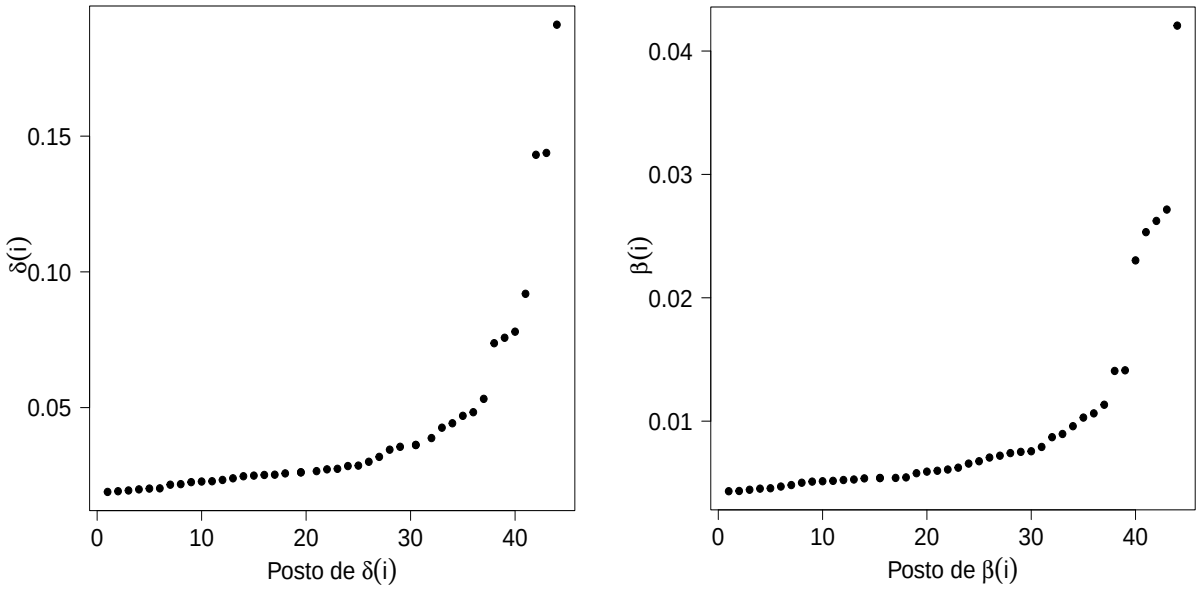
Na Figura 2 (a) as sequências que se afastam do padrão linear das demais são indicadas com ■ ao invés de • e nomeadas. A correlação amostral entre $\delta(i)$ e $\beta(i)$ obtida, desconsiderando-se as sequências KY817930.1 e KY785439.1 foi de 0,9439 (frente a 0,9099 onde todas as sequências são consideradas).



(a) (b)

Figura 2 – Diagramas de dispersão entre: (a) $\delta(i)$ e $\beta(i)$ e (b) as estatísticas de ordem de $\delta(i)$ e $\beta(i)$. Os valores de $\delta(i)$ e $\beta(i)$ referem-se as sequências brasileiras apresentadas na Tabela 11.

Com base nos valores de $\delta(i)$, a sequência KY558999.1 é a mais representativa (menor $\delta(i)$) e, a KY785439.1, a menos representativa (maior $\delta(i)$). Em outras palavras, $KY558999.1 = \underset{i \in \mathcal{C}^{BRA}}{\arg\min}\{\delta(i)\}$, $KY785439.1 = \underset{i \in \mathcal{C}^{BRA}}{\arg\max}\{\delta(i)\}$, onde $\mathcal{C}^{BRA}$ é a coleção das amostras brasileiras. Além disso, $\delta(i)$ assume valores no intervalo $[0,01894; 0,19106]$ com uma amplitude de 0,17212. No entanto, o terceiro quartil de $\delta(i)$ é igual a 0,04301, ou seja, a amplitude entre o terceiro quartil e o mínimo é igual a 0,02407, frente a 0,17212 entre o máximo e o mínimo. Portanto, existe grande concentração de sequências que possuem valores baixos de $\delta(i)$. Essa grande concentração de sequências com valores baixos de $\delta(i)$ (e também de $\beta(i)$) também é observado nos gráficos de pontos da Figura 3.



(a) (b)

Figura 3 – Diagramas de dispersão entre: (a) os postos de $\delta(i)$ e $\delta(i)$; (b) os postos de $\beta(i)$ e $\beta(i)$. Os valores de $\delta(i)$ e $\beta(i)$ referem-se as sequências brasileiras apresentadas na Tabela 11.

De modo a avaliar possíveis grupos entre as sequências brasileiras, um dendrograma (Figura 1 de (GARCÍA et al., 2018)), usando como matriz de distâncias D_1 cujas entradas são d_{max} entre todas as sequências brasileiras, ou seja, $D_1 = [d_{max}(x_{i,1}^{n_i}, x_{j,1}^{n_j}), \forall i, j = 1, ..., 44]$ e sendo $x_{i,1}^{n_i}$ a i-ésima sequência brasileira observada. Os agrupamentos obtidos neste dendrograma são apresentados na Tabela 12 e, algumas estatísticas descritivas destes, na Tabela 13.

Com base nas Tabelas 12 e 13, temos 4 (quatro) grupos de sequências, onde um dos grupos, D^{BRA}, possui 79,5 % das sequências. Além disso, temos que os grupos correspondem a valores quase que sequenciais de $\delta(i)$ (repare que existe pouca intersecção entre os valores mínimos e máximos de grupos vizinhos), ou seja, poder-se-ia definir esses grupos (agrupamentos) por meio da medida $\delta(i)$.

Uma outra informação interessante apresentada na Tabela 13 diz respeito ao coeficiente de variação de $\delta(i)$. Veja que, para todo o conjunto de sequências, o coeficiente de variação é de 0,86, ao passo que, após a formação dos grupos, a variabilidade dentro desses grupos foi inferior a 0,29, ou seja, uma diminuição substancial na variabilidade dentro dos grupos.

Tabela 12 – Agrupamentos (*clusters*) obtidos do dendrograma da Figura 1 de (GARCÍA et al., 2018) para as 44 sequências brasileiras (vide coluna 1 da Tabela 11).

Agrupamento	Sequências
A^{BRA}	KY559004.1, KY785439.1 e KY817930.1 .
B^{BRA}	KY014308.2, KY559001.1 e KY559014.1 .
C^{BRA}	KY559009.1, KY559010.1 e KY559031.1 .
D^{BRA}	KX197192.1, KY014296.2, KY014297.2, KY014301.2, KY014307.2, KY014313.2, KY014317.2, KY014320.2, KY558999.1, KY559003.1, KY559005.1, KY559006.1, KY559007.1, KY559011.1, KY559012.1, KY559013.1, KY559015.1, KY559017.1, KY559018.1, KY559019.1, KY559021.1, KY559023.1, KY559024.1, KY559027.1, KY559032.1, KY785410.1, KY785426.1, KY785427.1, KY785429.1, KY785433.1, KY785437.1, KY785450.1, KY785455.1, KY785456.1 e KY785479.1.

Tabela 13 – Estatísticas descritivas dos agrupamentos (*clusters*) indicados na Tabela 12. As informações disponíveis são: (a) quantidade de sequências no grupo (# grupo), (b) proporção de elementos no grupo (% # grupo), (c) $\delta(i)_{(1)}$: menor valor de $\delta(i)$ no grupo, (d) $\delta(i)_{(n)}$: maior valor de $\delta(i)$, (e) valor médio de $\delta(i)$ no grupo (Média de $\delta(i)$) e (f) coeficiente de variação de $\delta(i)$ no grupo (CV $\delta(i)$) .

Agrupamento	# grupo	% # grupo	$\delta(i)_{(1)}$	$\delta(i)_{(n)}$	Média de $\delta(i)$	CV $\delta(i)$
A^{BRA}	3	0,068	42	44	0,15934	0,17244
B^{BRA}	3	0,068	39	41	0,08187	0,10715
C^{BRA}	3	0,068	33	38	0,05651	0,27992
D^{BRA}	35	0,795	1	36	0,02789	0,28023
$\mathcal{C}^{BRA}$	44				0,04248	0,86381

Sequências norte-americanas Os valores de $\delta(i)$ e $\beta(i)$ para o universo das sequências norte-americanas, juntamente com a relação destas sequências, são apresentadas na Tabela 14, adaptada de (GARCÍA et al., 2018). Os valores dos postos de $\delta(i)$ e $\beta(i)$ são apresentados, nas colunas $\delta_{(k)}$ e $\beta_{(k)}$, respectivamente. Assim como nas sequências brasileiras, por exemplo, o valor de $\delta(i)$ associado a amostra KY075936.1 ($\delta = 0,01045$) corresponde a $\delta(i)_{(2)}$, ou seja, o segundo menor valor de δ e, o valor de $\beta(i)$ atrelado a mesma amostra KY075936.1 ($\beta = 0,00229$) corresponde a $\beta(i)_{(7)}$, o seja, o sétimo menor valor de β. As sequências estão ordenadas, de acordo com a magnitude de $\delta(i)$, em ordem crescente. Na tabela 15 expomos também algumas estatísticas descritivas, calculadas a partir dos valores obtidos de $\delta(i)$ e $\beta(i)$ (terceira e quinta colunas).

Novamente, como ocorreu com as sequências brasileiras, existe certa concordância entre: os valores calculados de $\delta(i)$ e $\beta(i)$ (vide Figura 4) e, os postos de $\delta(i)$ e $\beta(i)$

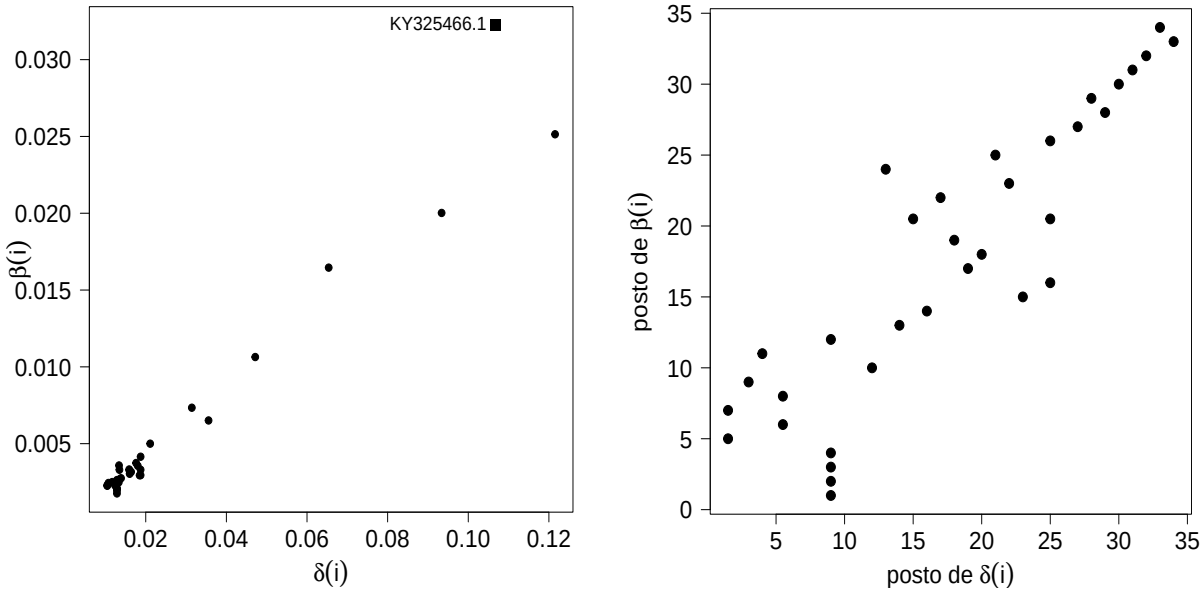
Tabela 14 – Da esquerda para a direita (a) sequências norte-americanas de vírus Zika, (b) valor de $\delta(i)$, (c) posto de $\delta(i)$ (Definição 3 (i)) (d) valor de $\beta(i)$ (Definição 3 (ii)) e (e) posto de $\beta(i)$. Sequências ordenadas em ordem crescente, de acordo com o valor de $\delta(i)$.

Sequência	δ	$\delta_{(k)}$	β	$\beta_{(k)}$
KX922706.1	0,01045	1,5	0,00226	5
KY075936.1	0,01045	1,5	0,00229	7
KX922707.1	0,01075	3	0,00244	9
KX922705.1	0,01168	4	0,00250	11
KX922704.1	0,01241	5	0,00227	6
KY014325.2	0,01241	6	0,00232	8
KY325472.1	0,01286	7	0,00209	4
KY014295.2	0,01286	8	0,00206	3
KY325469.1	0,01286	9	0,00195	2
KX832731.1	0,01286	10	0,00263	12
KY325468.1	0,01286	11	0,00176	1
KY014316.2	0,01331	12	0,00248	10
KY325465.1	0,01337	13	0,00358	24
KY325467.1	0,01345	14	0,00267	13
KY325479.1	0,01347	15	0,00330	20,5
KY325473.1	0,01387	16	0,00275	14
KY075934.1	0,01589	17	0,00332	22
KY785445.1	0,01596	18	0,00325	19
KY325476.1	0,01601	19	0,00304	17
KY325477.1	0,01637	20	0,00316	18
KY075932.1	0,01761	21	0,00374	25
KY075935.1	0,01802	22	0,00356	23
KY325464.1	0,01848	23	0,00294	15
KX842449.2	0,01873	24	0,00295	16
KY075933.1	0,01873	25	0,00415	26
KX922703.1	0,01873	26	0,00330	20,5
KY785459.1	0,02110	27	0,00500	27
KY785412.1	0,03143	28	0,00734	29
KY014298.1	0,03558	29	0,00651	28
KY014326.1	0,04716	30	0,01064	30
KY325471.1	0,06538	31	0,01646	31
KY785474.1	0,09339	32	0,02002	32
KY325466.1	0,10675	33	0,03223	34
KY785457.1	0,12153	34	0,02514	33

Fonte: Adaptação da Tabela 6 de (GARCÍA et al., 2018).

(vide Figura 5). A correlação linear amostral entre $\delta(i)$ e $\beta(i)$ calculada foi de 0,97699, ou seja, superior ao resultado obtido para as sequências brasileiras (0,9099). Já a correlação amostral obtida entre os postos de $\delta(i)$ e $\beta(i)$ foi de 0,89306 (inferior ao obtido para o caso das sequências brasileiras, de 0,9385). Portanto, de forma análoga ao que ocorreu com as sequências brasileiras, para as sequências norte-americanas também existe uma associação linear positiva entre δ e β e, também, entre os postos destas quantidades.

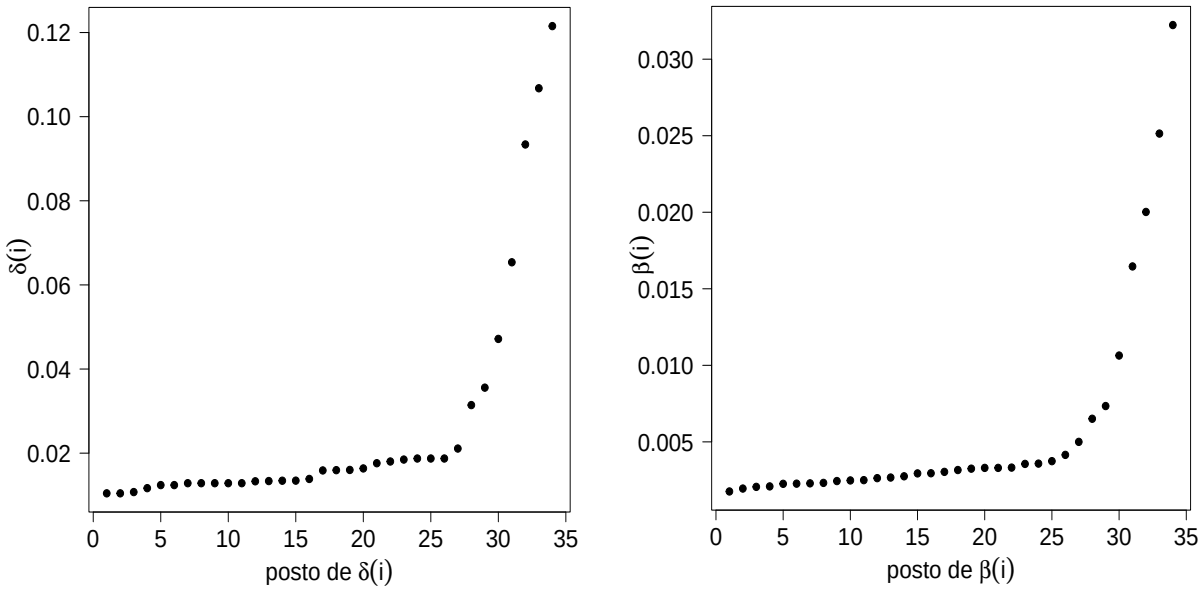
Na Figura 4 (a), como no caso das sequências brasileiras, a sequência que se afastou do padrão linear das demais é representada com ▪ ao invés de • e indicada (no caso, a KY325466.1). A correlação amostral entre $\delta(i)$ e $\beta(i)$ obtida, desconsiderando-se



(a) (b)

Figura 4 – Diagramas de dispersão entre: (a) $\delta(i)$ e $\beta(i)$ e (b) os postos de $\delta(i)$ e $\beta(i)$. Os valores de $\delta(i)$ e $\beta(i)$ referem-se as sequências norte-americanas apresentadas na Tabela 14.

KY325466.1 foi de 0,9922 (frente a 0,9770 onde todas as sequências são consideradas).



(a) (b)

Figura 5 – Diagramas de dispersão entre: (a) os postos de $\delta(i)$ e $\delta(i)$; (b) os postos de $\beta(i)$ e $\beta(i)$. Os valores de $\delta(i)$ e $\beta(i)$ referem-se as sequências norte-americanas apresentadas na Tabela 14.

O intervalo de variação de $\delta(i)$ é [0,01045; 0,12153] (no caso das sequências brasileiras, lembremos que era [0,01894; 0,19106]) e, com uma amplitude de 0,11108 (no caso brasileiro, 0,17212). Em comparação com as sequências brasileiras, as norte-americanas

têm $\delta(i)$ e $\beta(i)$ mais fortemente correlacionados e, além disso, valores de $\delta(i)$ mais próximos, uma vez que tem-se uma amplitude amostral menor. O mínimo e o terceiro quartil de δ^{EUA} (vide Tabela 15) são 0,01045 e 0,01873, respectivamente. Assim, a diferença entre o terceiro quartil e o mínimo é 0,00828 (frente a amplitude total de 0,11108). Portanto, a maioria das amostras norte-americanas possui valores de δ^{EUA} baixos. Essa última afirmação é também corroborada pelos gráficos da Figura 5.

Tabela 15 – Estatísticas descritivas de δ e β obtidos das sequências brasileiras e norte americanas.

Estatística	δ^{BRA}	δ^{EUA}	β^{BRA}	β^{EUA}
Mínimo	0,01894	0,01045	0,00432	0,00176
Q1	0,02323	0,01286	0,00521	0,00245
Mediana	0,02735	0,01593	0,00616	0,00310
Média	0,04248	0,02638	0,00930	0,00577
Q3	0,04301	0,01873	0,00912	0,00405
Máximo	0,19106	0,12153	0,04206	0,03223
DP	0,03670	0,02807	0,00778	0,00709
CV	0,86381	1,06437	0,83750	1,22931

Q1: primeiro quartil, Q3: terceiro quartil, DP: desvio padrão, CV: coeficiente de variação.

Tomando os valores de $\delta(i)$, as sequência KX922706.1 e KY075936.1 são as mais representativas (menores $\delta(i)$) e, a KY785457.1, a menos (maior $\delta(i)$). Em outras palavras,

$$KX922706.1(\text{ou KY075936.1}) = \arg\min_{i \in \mathcal{C}^{EUA}} \{\delta(i)\},$$
$$KY785457.1 = \arg\max_{i \in \mathcal{C}^{EUA}} \{\delta(i)\},$$

onde $\mathcal{C}^{EUA}$ é a coleção das amostras norte-americanas. Portanto, como apresentado em (GARCÍA et al., 2018) as amostras $KX922706.1$ e $KY075936.1$ deveriam ser as amostras escolhidas, em caso de necessidade de expressar-se o comportamento padrão das cepas norte-americanas. Além disso, a sequência KY785457.1 é a que apresenta um comportamento mais divergente, uma vez que seu valor $\delta(i)$ é o maior possível no universo considerado.

Um dendrograma também foi construído tendo por base uma matriz de distâncias D_2 tendo como entradas as quantidades d_{max} entre as sequências norte-americanas, i.e.,

$$D_2 = [d_{max}(x_{i,1}^{n_i}, x_{j,1}^{n_j}), \forall i, j = 1, ..., 34],$$

e $x_{i,1}^{n_i}$ sendo a i-ésima sequência norte-americana medida.

Esse dendrograma é apresentado na Figura 2 de (GARCÍA et al., 2018) onde são identificados claramente 4 (quatro) grupos, discriminados na Tabela 16. Também, algumas estatísticas descritivas destes grupos são apresentadas na Tabela 17.

Um dos grupos formados possui praticamente 80% das sequências. Além disso, temos que os grupos correspondem a valores quase, assim como também aconteceu com as sequências brasileiras, que sequenciais de $\delta(i)$ (repare que não existe intersecção entre os valores mínimos e máximos de grupos vizinhos), ou seja, poder-se-ia construir esses grupos (agrupamentos) usando a medida $\delta(i)$. Por fim, como aconteceu com o grupo das sequências brasileiras, a variabilidade de $\delta(i)$ dos grupos formados foi bastante inferior a variabilidade total de $\delta(i)$ (Tabela 17).

Tabela 16 – Agrupamentos (*clusters*) obtidos do dendrograma da Figura 2 de (GARCÍA et al., 2018) para as 34 sequências norte-americanas (vide coluna 1 da Tabela 14).

Agrupamento	Sequências
A^{EUA}	KY325466.1, KY785457.1, KY785474.1 .
B^{EUA}	KY325471.1 .
C^{EUA}	KY014298.1, KY014326.1, KY785412.1 .
D^{EUA}	KX832731.1, KX842449.2, KX922703.1, KX922704.1, KX922705.1, KX922706.1, KX922707.1, KY014295.2, KY014316.2, KY014325.2, KY075932.1, KY075933.1, KY075934.1, KY075935.1, KY075936.1, KY325464.1, KY325465.1, KY325467.1, KY325468.1, KY325469.1, KY325472.1, KY325473.1, KY325476.1, KY325477.1, KY325479.1, KY785445.1, KY785459.1 .

Tabela 17 – Estatísticas descritivas dos agrupamentos (*clusters*) indicados na Tabela 16. As informações disponíveis são: (a) quantidade de sequências no grupo (# grupo), (b) proporção de elementos no grupo (% # grupo), (c) $\delta(i)_{(1)}$: menor valor de $\delta(i)$ no grupo, (d) $\delta(i)_{(n)}$: maior valor de $\delta(i)$, (e) valor médio de $\delta(i)$ no grupo (Média de $\delta(i)$) e (f) coeficiente de variação de $\delta(i)$ no grupo (CV $\delta(i)$) .

Agrupamento	# grupo	% # grupo	$\delta(i)_{(1)}$	$\delta(i)_{(n)}$	Média de $\delta(i)$	CV $\delta(i)$
A^{EUA}	3	0,088	32	34	0,10722	0,13128
B^{EUA}	1	0,029	31	31	0,06538	-
C^{EUA}	3	0,088	28	30	0,03806	0,21421
D^{EUA}	27	0,794	1	27	0,01465	0,20153
$\mathcal{C}^{EUA}$	34				0,02638	1,06437

Todas as sequências Nessa etapa do trabalho, de modo a facilitar a identificação do país de origem da sequência analisada, incorporamos ao nome de cada uma das sequências, a abreviação do nome desse país. As abreviações usadas são as mesmas usadas no parágrafo *dados genômicos* em 2.4.1.1. A relação de todas as sequências é apresentada na Tabela 18.

As Tabelas 19 e 20 apresentam os valores de $\delta(i)$ de cada uma das 153 sequências, relacionadas na Tabela 18. Algumas estatísticas descritivas também são apresentadas (Tabela 21). Observamos valores de $\delta(i)$ no intervalo [0,01456; 0,19644] com amplitude de 0,18188, mediana de 0,02064 e terceiro quartil de 0,03697. Portanto, a amplitude dos 75% valores de $\delta(i)$ menores é de 0,02241 (frente a amplitude total de 0,18188), ou seja,

Tabela 18 – Sequências do vírus Zika por país: BRA (Brasil), USA (Estados Unidos), DOM (República Dominicana), MEX (México), HND (Honduras), NIC (Nicarágua), JAM (Jamaica), COL (Colômbia), PRI (Porto Rico), VEN (Venezuela), CUB (Cuba) e MTQ (Martinica). A coluna 2, fornece a quantidade de sequências de cada país.

País	# Seq	Sequências
BRA	44	KY559018.1, KY559013.1, KY559010.1, KY785437.1, KY014320.2, KY559021.1, KY559012.1, KY785410.1, KY785427.1, KY014296.2, KY559017.1, KY014308.2, KY014301.2, KY785439.1, KY785456.1, KY559001.1, KY559032.1, KY559005.1, KY014297.2, KY785429.1, KY559014.1, KY785426.1, KY559023.1, KY559004.1, KY785450.1, KY559031.1, KY559015.1, KY817930.1, KY014313.2, KX197192.1, KY014307.2, KY014317.2, KY559009.1, KY559024.1, KY785479.1, KY785455.1, KY559019.1, KY559003.1, KY785433.1, KY559011.1, KY559006.1, KY559027.1, KY559007.1, KY558999.1
USA	34	KY075933.1, KY785457.1, KY014316.2, KY075936.1, KY014326.1, KY014325.2, KY785445.1, KX922705.1, KY075932.1, KX922706.1, KY325468.1, KY325465.1, KY075935.1, KY325476.1, KX922707.1, KY325464.1, KY325477.1, KY785474.1, KX922703.1, KY325469.1, KY325479.1, KX832731.1, KY014298.1, KX842449.2, KY325472.1, KY325473.1, KY325471.1, KX922704.1, KY785459.1, KY075934.1, KY325466.1, KY325467.1, KY014295.2, KY785412.1
DOM	23	KY785484.1, KY785465.1, KY014304.2, KY785475.1, KY014321.2, KY785470.1, KY785435.1, KY014300.2, KY785453.1, KY785463.1, KY785441.1, KY785476.1, KY014305.2, KY785447.1, KY785449.1, KY785415.1, KY785420.1, KY785413.1, KY014314.2, KY014318.3, KY014302.3, KY785423.1, KY014303.2
MEX	19	MF801395.1, MF801410.1, MF801404.1, MF801403.1, MF801406.1, MF801409.1, MF801407.1, MF801413.1, MF801398.1, MF801414.1, MF801423.1, MF801408.1, MF801396.1, MF801412.1, MF801420.1, MF801402.1, MF801418.1, MF801417.1, MF801391.1
HND	13	KY014306.2, KY785452.1, KY014310.2, KY785461.1, KY785448.1, KY014327.2, KY014319.2, KY014312.2, KY785444.1, KY014315.2, KY785442.1, KY785414.1, KY785418.1
NIC	7	MF434517.1, MF434522.1, MF801426.1, MF434516.1, MF434520.1, MF434518.1, MF434521.1
JAM	4	KY785430.1, KY785424.1, KY785419.1, KY785432.1
COL	3	KY785469.1, KY785466.1, KY785417.1
PRI	2	KY785464.1, KY785462.1
VEN	2	KX893855.1, KX702400.1
CUB	1	MF438286.1
MTQ	1	KY785451.1

Fonte: Adaptação da Tabela 1 de (GARCÍA et al., 2018) .

a distribuição dos $\delta(i)$, do mesmo jeito que aconteceu com as sequências brasileiras e norte-americanas, é bastante assimétrica à direita. Além disso, sendo

$$KY014318.3.DOM = \arg\min_{i\in\mathcal{C}} \{\delta(i)\} \quad \text{e}$$

$$KY785444.1.HND = \arg\max_{i\in\mathcal{C}} \{\delta(i)\}$$

onde $\mathcal{C}$ é a coleção de todas as sequências, temos que as sequências, a luz de $\delta(i)$, KY014318.3.DOM e KY785444.1.HND são as mais e menos representativas de todas as sequências, respectivamente.

Tabela 19 – 1 de 2: Sequências do vírus Zika (listadas na Tabela 18) ordenadas em ordem crescente, em relação ao valor de $\delta(i)$ e, $\delta_{(k)}$ é o posto de $\delta(i)$.

Sequência	$\delta(i)$	$\delta_{(k)}$	Sequência	$\delta(i)$	$\delta_{(k)}$
KY014318.3.DOM	0,01456	1	MF801391.1.MEX	0,01866	40
KY785435.1.DOM	0,01502	2	KY014310.2.HND	0,01867	41
KY785420.1.DOM	0,01543	3	MF801414.1.MEX	0,01878	42
KX922706.1.USA	0,01583	4	KY785441.1.DOM	0,01890	43
KY014302.3.DOM	0,01586	5	KY014307.2.BRA	0,01900	44
KY559005.1.BRA	0,01592	6	MF801413.1.MEX	0,01915	45
KY075934.1.USA	0,01601	7	KY559027.1.BRA	0,01922	46
KY014306.2.HND	0,01604	8	MF434516.1.NIC	0,01924	47
KX922707.1.USA	0,01621	9	MF801406.1.MEX	0,01936	48
KY785464.1.PRI	0,01628	10	KY014303.2.DOM	0,01938	49
KY559007.1.BRA	0,01634	11	KY014321.2.DOM	0,01942	50,5
KX832731.1.USA	0,01649	12	KY014312.2.HND	0,01942	50,5
MF801418.1.MEX	0,01660	13	MF801410.1.MEX	0,01943	52,5
KY075936.1.USA	0,01674	14	MF801417.1.MEX	0,01943	52,5
KY785450.1.BRA	0,01681	15	KY325473.1.USA	0,01949	54
KY559015.1.BRA	0,01689	16	KY785418.1.HND	0,01973	55,5
KY014319.2.HND	0,01690	17	KY014317.2.BRA	0,01973	55,5
KY558999.1.BRA	0,01698	18	KY785465.1.DOM	0,01980	57
MF801426.1.NIC	0,01714	19	KY325472.1.USA	0,01984	58
MF801403.1.MEX	0,01750	20	KY014295.2.USA	0,01984	59
MF801395.1.MEX	0,01758	21	KY325468.1.USA	0,01984	60
KY075932.1.USA	0,01761	22	KY014316.2.USA	0,01985	61
KY559013.1.BRA	0,01762	23	MF801398.1.MEX	0,01992	62
KY325465.1.USA	0,01777	24,5	KY325469.1.USA	0,01996	63,5
KY325479.1.USA	0,01777	24,5	MF434521.1.NIC	0,01996	63,5
MF801412.1.MEX	0,01778	26	KX702400.1.VEN	0,02001	65,5
KY785415.1.DOM	0,01814	27	KX893855.1.VEN	0,02001	65,5
KX922704.1.USA	0,01819	28,5	MF801402.1.MEX	0,02002	67
KY014325.2.USA	0,01819	28,5	KY075933.1.USA	0,02005	68
KY785455.1.BRA	0,01825	30	KY785410.1.BRA	0,02012	69
KY014297.2.BRA	0,01831	31	KY075935.1.USA	0,02017	70
MF801396.1.MEX	0,01832	32	KY785442.1.HND	0,02031	71
KY785476.1.DOM	0,01839	33	KY785484.1.DOM	0,02052	72
KY785445.1.USA	0,01841	34	KX842449.2.USA	0,02059	73,5
KY014296.2.BRA	0,01842	35,5	KX922703.1.USA	0,02059	73,5
KY014320.2.BRA	0,01842	35,5	KY785452.1.HND	0,02060	75
KX922705.1.USA	0,01851	37	KY785479.1.BRA	0,02062	76
KY014314.2.DOM	0,01859	38,5	KX197192.1.BRA	0,02064	77
KY785475.1.DOM	0,01859	38,5			

Fonte: Adaptação das Tabelas 8, 9 e 10 de (GARCÍA et al., 2018).

Tabela 20 – 2 de 2: Sequências do vírus Zika (listadas na Tabela 18) ordenadas em ordem crescente, em relação ao valor de $\delta(i)$ e, $\delta_{(k)}$ é o posto de $\delta(i)$.

Sequência	$\delta(i)$	$\delta_{(k)}$	Sequência	$\delta(i)$	$\delta_{(k)}$
KY014300.2.DOM	0,02074	78	KY559019.1.BRA	0,03848	116
KY785433.1.BRA	0,02102	79	KY559018.1.BRA	0,03849	117
KY014304.2.DOM	0,02112	80	KY014298.1.USA	0,03968	118
MF434522.1.NIC	0,02118	81	MF434518.1.NIC	0,03979	119
MF434517.1.NIC	0,02160	82	KY785414.1.HND	0,04050	120
MF801408.1.MEX	0,02172	83	KY014313.2.BRA	0,04116	121
KY325464.1.USA	0,02198	84	KY785424.1.JAM	0,04167	122
KY785419.1.JAM	0,02199	85,5	KY559031.1.BRA	0,04571	123
KY325476.1.USA	0,02199	85,5	KY785437.1.BRA	0,04642	124
KY325477.1.USA	0,02232	87	KY559032.1.BRA	0,05038	125
KY325467.1.USA	0,02251	88	KY559011.1.BRA	0,05433	126
KY014315.2.HND	0,02297	89	KY014326.1.USA	0,05496	127
KY785448.1.HND	0,02322	90,5	KY785430.1.JAM	0,05648	128
KY785427.1.BRA	0,02322	90,5	MF801404.1.MEX	0,05702	129
KY785426.1.BRA	0,02358	92	KY785413.1.DOM	0,05936	130
KY014301.2.BRA	0,02403	93	KY559010.1.BRA	0,06237	131
KY014327.2.HND	0,02423	94	KY325471.1.USA	0,06359	132
KY785469.1.COL	0,02438	95	KY785451.1.MTQ	0,07057	133
KY785423.1.DOM	0,02447	96	KY014308.2.BRA	0,07579	134
KY785456.1.BRA	0,02500	97	KY559009.1.BRA	0,07588	135
KY785453.1.DOM	0,02516	98	MF801407.1.MEX	0,07824	136
KY559012.1.BRA	0,02570	99	KY559001.1.BRA	0,08650	137
KY785459.1.USA	0,02607	100	KY559014.1.BRA	0,09922	138
KY014305.2.DOM	0,02613	101	KY785474.1.USA	0,10512	139
KY559021.1.BRA	0,02614	102	KY785417.1.COL	0,10523	140
MF438286.1.CUB	0,02638	103	KY325466.1.USA	0,10613	141
KY785466.1.COL	0,02674	104	KY785447.1.DOM	0,11085	142
KY785429.1.BRA	0,02902	105	MF801420.1.MEX	0,11465	143
KY785462.1.PRI	0,02980	106	KY785461.1.HND	0,11535	144
KY559024.1.BRA	0,03148	107,5	KY785457.1.USA	0,11971	145
KY559023.1.BRA	0,03148	107,5	KY785432.1.JAM	0,12871	146
KY559017.1.BRA	0,03161	109	MF434520.1.NIC	0,13332	147
MF801423.1.MEX	0,03262	110,5	KY817930.1.BRA	0,14631	148
KY559003.1.BRA	0,03262	110,5	KY559004.1.BRA	0,15752	149
KY559006.1.BRA	0,03409	112	MF801409.1.MEX	0,16916	150
KY785470.1.DOM	0,03432	113	KY785449.1.DOM	0,17585	151
KY785412.1.USA	0,03673	114	KY785439.1.BRA	0,19191	152
KY785463.1.DOM	0,03697	115	KY785444.1.HND	0,19644	153

Fonte: Adaptação das Tabelas 8, 9 e 10 de (GARCÍA et al., 2018).

Tabela 21 – Estatísticas descritivas de δ obtidas de todas as sequências (Vide Tabelas 19 e 20).

Estatística	δ
Mínimo	0,01045
Q1	0,01286
Mediana	0,01593
Média	0,02638
Q3	0,01873
Máximo	0,12153
DP	0,02807
CV	1,06437

Q1: primeiro quartil, Q3: terceiro quartil, DP: desvio padrão, CV: coeficiente de variação.

A Figura 3 de (GARCÍA et al., 2018) exibe um dendrograma, cuja matriz de distâncias possui como entradas d_{max} calculadas para todas as sequências indicadas na Tabela 18, com função de ligação média. Os 4 (quatro) grupos obtidos e algumas de suas estatísticas descritivas são apresentados nas Tabelas 22 e 23, respectivamente.

Tabela 23 – Estatísticas descritivas dos agrupamentos (*clusters*) indicados na Tabela 22. As informações disponíveis são: (a) quantidade de sequências no grupo (# grupo), (b) proporção de elementos no grupo (% # grupo), (c) menor valor de $\delta(i)$ no grupo, (d) maior valor de $\delta(i)$, (e) valor médio de $\delta(i)$ no grupo (Média de $\delta(i)$) e (f) coeficiente de variação de $\delta(i)$ no grupo (CV $\delta(i)$) .

Agrupamento	# grupo	% # grupo	$\delta(i)_{(1)}$	$\delta(i)_{(n)}$	Média de $\delta(i)$	CV $\delta(i)$
A^{geral}	7	0,046	147	153	0,16722	0,13859
B^{geral}	16	0,105	129	146	0,09454	0,23268
C^{geral}	14	0,092	114	131	0,04762	0,18123
D^{geral}	116	0,758	1	119	0,02121	0,24098
$\mathcal{C}^{GERAL}$	153				0,03797	0,99382

Com base nas Tabelas 22 e 23, temos 4 (quatro) grupos de sequências, onde um dos grupos, D^{Geral}, possui 75,8 % das sequências. Além disso, temos que os grupos correspondem a valores quase que sequenciais de $\delta(i)$ (repare que existe pouca intersecção entre os valores mínimos e máximos de grupos vizinhos), ou seja, poder-se-ia definir esses grupos (agrupamentos) por meio da noção $\delta(i)$.

Uma outra informação interessante apresentada na Tabela 23 diz respeito ao coeficiente de variação de $\delta(i)$. Veja que, para todo o conjunto de sequências, o coeficiente de variação é de 0,99, ao passo que, após a formação dos grupos, a variabilidade dentro desses grupos foi inferior a 0,25, ou seja, uma diminuição substancial na variabilidade dentro dos grupos.

Tabela 22 – Agrupamentos (*clusters*) obtidos do dendrograma da Figura 3 de (GARCÍA et al., 2018), onde todas as sequências indicadas na Tabela 18 são consideradas.

Agrupamento	Sequências
A^{geral}	KY559004.1.BRA, KY785439.1.BRA, KY785444.1.HND, KY785449.1.DOM, KY817930.1.BRA, MF434520.1.NIC e MF801409.1.MEX .
B^{geral}	KY014308.2.BRA, KY325466.1.USA, KY325471.1.USA, KY559001.1.BRA, KY559009.1.BRA, KY559014.1.BRA, KY785417.1.COL, KY785432.1.JAM, KY785447.1.DOM, KY785451.1.MTQ, KY785457.1.USA, KY785461.1.HND, KY785474.1.USA, MF801404.1.MEX, MF801407.1.MEX e MF801420.1.MEX .
C^{geral}	KY014298.1.USA, KY014313.2.BRA, KY014326.1.USA, KY559010.1.BRA, KY559011.1.BRA, KY559031.1.BRA, KY559032.1.BRA, KY785412.1.USA, KY785413.1.DOM, KY785414.1.HND, KY785424.1.JAM, KY785430.1.JAM, KY785437.1.BRA e KY785463.1.DOM .
D^{geral}	KX197192.1.BRA, KX702400.1.VEN, KX832731.1.USA, KX842449.2.USA, KX893855.1.VEN, KX922703.1.USA, KX922704.1.USA, KX922705.1.USA, KX922706.1.USA, KX922707.1.USA, KY014295.2.USA, KY014296.2.BRA, KY014297.2.BRA, KY014300.2.DOM, KY014301.2.BRA, KY014302.3.DOM, KY014303.2.DOM, KY014304.2.DOM, KY014305.2.DOM, KY014306.2.HND, KY014307.2.BRA, KY014310.2.HND, KY014312.2.HND, KY014314.2.DOM, KY014315.2.HND, KY014316.2.USA, KY014317.2.BRA, KY014318.3.DOM, KY014319.2.HND, KY014320.2.BRA, KY014321.2.DOM, KY014325.2.USA, KY014327.2.HND, KY075932.1.USA, KY075933.1.USA, KY075934.1.USA, KY075935.1.USA, KY075936.1.USA, KY325464.1.USA, KY325465.1.USA, KY325467.1.USA, KY325468.1.USA, KY325469.1.USA, KY325472.1.USA, KY325473.1.USA, KY325476.1.USA, KY325477.1.USA, KY325479.1.USA, KY558999.1.BRA, KY559003.1.BRA, KY559005.1.BRA, KY559006.1.BRA, KY559007.1.BRA, KY559012.1.BRA, KY559013.1.BRA, KY559015.1.BRA, KY559017.1.BRA, KY559018.1.BRA, KY559019.1.BRA, KY559021.1.BRA, KY559023.1.BRA, KY559024.1.BRA, KY559027.1.BRA, KY785410.1.BRA, KY785415.1.DOM, KY785418.1.HND, KY785419.1.JAM, KY785420.1.DOM, KY785423.1.DOM, KY785426.1.BRA, KY785427.1.BRA, KY785429.1.BRA, KY785433.1.BRA, KY785435.1.DOM, KY785441.1.DOM, KY785442.1.HND, KY785445.1.USA, KY785448.1.HND, KY785450.1.BRA, KY785452.1.HND, KY785453.1.DOM, KY785455.1.BRA, KY785456.1.BRA, KY785459.1.USA, KY785462.1.PRI, KY785464.1.PRI, KY785465.1.DOM, KY785466.1.COL, KY785469.1.COL, KY785470.1.DOM, KY785475.1.DOM, KY785476.1.DOM, KY785479.1.BRA, KY785484.1.DOM, MF434516.1.NIC, MF434517.1.NIC, MF434518.1.NIC, MF434521.1.NIC, MF434522.1.NIC, MF438286.1.CUB, MF801391.1.MEX, MF801395.1.MEX, MF801396.1.MEX, MF801398.1.MEX, MF801402.1.MEX, MF801403.1.MEX, MF801406.1.MEX, MF801408.1.MEX, MF801410.1.MEX, MF801412.1.MEX, MF801413.1.MEX, MF801414.1.MEX, MF801417.1.MEX, MF801418.1.MEX, MF801423.1.MEX e MF801426.1.NIC .

Conclusão Neste estudo de caso, usamos o $d_{\max}$ (Definição 2) para estabelecermos proximidade, no que diz respeito a sua lei probabilística de formação, entre as sequências DNA do vírus da Zika. Tendo por base as propriedades teóricas de $d_{\max}$ e, pelo fato do menor tamanho amostral $n = 10807$ ser grande o suficiente, podemos afirmar que todas as sequências são geradas por uma mesma lei, uma vez que $d_{\max} < 1$ em todos os casos. Esse resultado não foi surpreendente, uma vez que todas as sequências são cadeias de DNA do vírus Zika. Independente disso, nosso resultado foi além, uma vez que, usando $\delta(\cdot)$ somos capazes de apontar quais sequências melhor representam o conjunto de sequências.

Nos focamos em três cenários distintos: (a) sequências do Brasil, (b) sequências

dos EUA e (c) o conjunto de 153 sequências mostradas na Tabela 18, vindas de 12 países distintos. Tendo por base esses cenários, identificamos aquelas sequências que são *localmente* mais representativas e, ao mesmo tempo, as que são menos. No cenário (a), a amostra mais representativa é a KY558999.1 ($\delta = 0,01894$), i.e., se fossemos escolher uma sequência que melhor representasse as sequências de DNA do vírus da Zika analisados para o Brasil, a escolhida seria a KY558999.1. Já no cenário (b), as amostras mais representativas são a KX922706.1 e KY075936.1 ($\delta = 0,01045$). Finalmente, no (c), a mais representativa é a KY014318.3DOM com seu $\delta = 0,01456$.

Adicionalmente, usando $\delta(\cdot)$, pudemos identificar quais as sequências que melhor representam a lei geradora dominante. A título de exemplo, as sequências KY014318.3.DOM, KY785435.1.DOM e KY785420.1.DOM as três sequências mais representativas quando consideramos todas as sequências juntas, ou seja, aquelas associadas aos três menores valores de δ. Sob a hipótese da existência de uma lei majoritária e se ao menos 50% de todas as sequências são formadas por esta lei, as três sequências citadas seguem essa lei majoritária.

Finalmente, neste estudo, fizemos uso de uma variedade de critérios consistentes provenientes dos processos estocásticos, que nos possibilitaram identificar as sequências ditas padrão (e outras que são raras ou menos representativas). Mostramos, com essa análise, como ferramentas de processos estocásticos podem ser úteis na tarefa de classificação de sequências de DNA.

2.4.2 Classificação do vírus da dengue tipo 1 autóctones que circularam no Japão em 2014

Apresentamos, adiante, uma descrição do trabalho (CORDEIRO et al., 2019a) (vide apêndice) já publicado, e que também é parte integrante da minha contribuição para o tema. Nesse estudo, fizemos a classificação, através da representatividade de elementos de um conjunto de sequências genômicas completas do Vírus da Dengue Tipo 1 (DENV-1) correspondentes ao surto de Dengue ocorrido no Japão em 2014, i.e., correspondem a casos autóctones. Essas sequências vieram de quatro prefeituras [11] japonesas, a saber: Chiba, Hyogo, Shizuoka e Tóquio. Esse conjunto foi considerado como sendo formado por amostras independentes de processos Markovianos de ordem e alfabeto finitos. Além disso, sob a hipótese de existência de uma lei estocástica dominante, no processo de geração das amostras, que ocorre em ao menos 50% das amostras do conjunto, identificamos as sequências governadas por essa lei (detalhes em (FERNÁNDEZ et al., 2019) e (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2018a)). A classificação dessas sequências foi feita com base no cálculo, entre as sequências, de d_{max}, ou seja, o máximo de d_s em todos os

[11] O Japão está subdividido em 47 prefeituras. As prefeituras são maiores que os organismos governamentais cidades, vilas e aldeias. Seriam, em termos brasileiros, equivalentes aos Estados.

estados $s \in \mathcal{S}$ considerados. Também, d_{max} tende a zero quando comparamos sequências com leis estocásticas formadoras idênticas e, diverge (ou seja, tende a ∞) quando as leis formadoras são diferentes (vide teorema 4). Nesse estudo, concluímos que as ordens de representatividade dessas sequências, da maior para a menor, de acordo com o local de origem das sequências é a seguinte: (i) Tóquio, (ii) Chiba, (iii) Hiogo, (iv) Shizuoka. Por fim, quando comparamos as sequências originadas no Japão com seus equivalentes asiáticos contemporâneos, encontramos que a sequência japonesa menos representativa (no caso a proveniente de Shizuoka) está consideravelmente longe (de acordo com uma análise de agrupamento feita com base numa matriz de distâncias cujas entradas são as quantidades d_{max}) do grupo formado pelas demais sequências vindas do Japão. Esse último comportamento sugere evidências que o surto ocorrido no Japão de 2014 poderia ter sido produzido por mais de um tipo de DENV-1.

2.4.2.1 Introdução

Neste estudo, fizemos a análise de 6 (seis) cepas de DENV-1 autóctones que foram isoladas de pacientes durante o surto de 2014 ocorrido no Japão (ver (TAJIMA et al., 2016)). O nosso objetivo com esse estudo é o de identificar as sequências mais e a menos representativas do conjunto considerado. Tal vírus é transmitido para seres humanos por mosquitos *Aedes* infectados. Ademais, até o momento, existem 5 (cinco) tipos do vírus da Dengue (MUSTAFA et al., 2015). Pessoas que já contraíram dengue sofrem o risco de, em contraindo novamente essa doença, desenvolverem formas graves da Dengue, como a Dengue Hemorrágica que é potencialmente fatal. Nos últimos anos, existe no mercado uma vacina para tal doença que costuma ser indicada para indivíduos que já contraíram dengue anteriormente. Desde a Segunda Guerra Mundial, os casos de Dengue identificados no Japão foram importados, no entanto, entre 2013 e 2014, mais de 160 casos autóctones foram identificados naquele país. Tal surto demandou uma investigação criteriosa da natureza desse surto e, dessa investigação chegou-se a conclusão que a contaminação foi causada pelo DENV-1. Os achados de (TAJIMA et al., 2016) sugerem que houve ao menos duas epidemias autóctones independentes no Japão em 2014 causadas pelo DENV-1.

Nosso principal objetivo, neste estudo, foi o de classificar as amostras do vírus da Dengue com relação a sua representatividade no grupo, ou seja, sabemos que como as amostras provém de pessoas distintas e, por isso, estão sujeitas a possuir construções genômicas distintas. Desejamos identificar as amostras "mais representativas", ou seja, aquelas que são as mais similares a todas as outras (em termos de lei de formação), do grupo. Além disso, queremos identificar a mais diferente, sob o mesmo critério, do grupo. Para chegar a esse objetivo, consideramos cada sequência como sendo uma amostra de um processo estocástico Markoviano e: (i) medimos as distâncias entre essas sequências usando a noção d_{max} (ver Definição 2) e, (ii) aplicamos um método robusto de classificação, introduzido em (FERNÁNDEZ et al., 2019) e usando a quantidade $\delta(i)$ (ver Definição 3

(i)) pudemos identificar a amostra mais representativa e, além disso, classificar todas as amostras em ordem de representatividade.

A questão do estabelecimento de proximidade entre sequências genômicas tem despertado o interesse de diversas áreas e com diferentes objetivos. A título de exemplo, (ZENG et al., 2005), (KWOK et al., 2012) e (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2018b), exploram alguma noção de similaridade entre cepas.

Um ponto diferente desse trabalho é que aqui a noção usada para o estabelecimento da proximidade entre sequências genômicas é uma métrica, no sentido matemático da palavra (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2018a). Além disso, essa métrica é também estatisticamente consistente [12].

O segundo aspecto e, talvez o mais relevante, seja o uso de um classificador baseado nessa métrica entre as sequências (veja (FERNÁNDEZ et al., 2019)). Esse classificador poderia ser usado futuramente na obtenção de representantes padrão para estruturas genômicas (CORDEIRO et al., 2019a) (vide apêndice).

2.4.2.2 Conjunto de dados e resultados

Inicialmente, apresentamos uma descrição, a fonte e a estrutura dos dados utilizados. Num segundo momento, calculamos as quantidades d_{max} e $\delta(i)$ entre as sequências, de modo a classificá-las por representatividade.

2.4.2.2.1 Vírus da Dengue Tipo 1

As sequências completas usadas nesse estudo estão disponíveis em <http://www.ncbi.nlm.nih.gov/> (NCBI - National Center for Biotechnology Information) e foram sequenciadas e estudadas pela primeira vez em (TAJIMA et al., 2016). As sequências são descritas na Tabela 24.

Tabela 24 – Sequências completas do vírus da dengue tipo 1. As colunas da esquerda para a direita representam: (1) o ID do paciente de onde vem a sequência, (2) o número de acesso a base NCBI, (3) a identificação da sequência/cepa, e (4) o possível local de contaminação do paciente. Fonte: Adaptado de (CORDEIRO et al., 2019a) (vide apêndice).

ID do paciente	Número acesso	Cepa	Área de contaminação
14-149J	LC011947	D1/Hu/Tokyo/NIID149/2014	No trem de Yotsuya-Shinjuku, Tóquio
14-111J	LC011946	D1/Hu/Tokyo/NIID111/2014	Próximo ao parque Yoyogi, Tóquio
14-181J	LC011949	D1/Hu/Shizuoka/NIID181/2014	Shizuoka?
14-188J	LC016760	D1/Hu/Hyogo/NIID188/2014	Hyogo? Malásia?
14-153J	LC011948	D1/Hu/Chiba/NIID153/2014	Chiba
14-100J	LC011945	D1/Hu/Saitama/NIID100/2014	Parque Yoyogi, Tóquio

[12] Ela recupera, para um n suficientemente grande, a classe de modelos correta, eventualmente quase certamente, desde que uma das classes de modelo candidatas esteja certa.

O possível epicentro da epidemia de dengue ocorrida em 2014 foi o Parque Yoyogi, em Tóquio. Uma parte das sequências provém de pacientes que passaram por lá ou por suas imediações. Informações detalhadas dos pacientes listados na Tabela 24 são fornecidas em (TAJIMA et al., 2016). Na última coluna da Tabela 24 informamos o local onde suspeita-se que tenha ocorrida a contaminação de cada um desses 6 pacientes. Alguns detalhes adicionais sobre os pacientes são dados em (TAJIMA et al., 2016).

De modo a ilustrar a estrutura dos dados, considere um pedaço inicial[13] da sequência com número de acesso LC011949 (cepa D1/Hu/Shizuoka/NIID181/2014),

$$...gagcagatct\ ctgatgaaca\ accaacggaa\ aaagacggct\ cgaccgtctt\ tcaatatgct...\,,$$

onde o alfabeto é $A = \{a, c, g, t\}$ com cardinalidade $|A| = 4$ com elementos: adenina (a), citosina (c), guanina (g) e timina (t). Todas as 6 sequências possuem em torno de 10700 elementos.

Inicialmente, para cada par de sequências, calculamos as métricas d_s, onde s é um estado do espaço de estados $\mathcal{S}$ e, também, d_{max}. Num segundo momento, de posse desses valores, procedemos com a classificação das sequências. Como de costume, em se tratando de dados genômicos, os elementos do alfabeto A são organizados em triplas e, assim, podemos escolher memórias múltiplos de 3, ou seja, $o = 3, 6, 9, ...$ Portanto, o espaço de estados $\mathcal{S}$ é composto pela concatenação de o elementos de A ($\mathcal{S} = A^o$). No nosso caso, o tamanho das sequências é de aproximadamente 10700, o que nos fornece uma memória recomendada de $o < \lfloor \log_{|A|}(10700) \rfloor - 1 = 7$, onde $\lfloor x \rfloor$ é o maior inteiro menor ou igual a x. Então, podemos fazer uso de uma memória 3 ou 6, e, por questões de simplicidade, optamos por uma memória $o = 3$.

2.4.2.3 Similaridade entre as sequências genômicas

Como o nosso interesse é uma comparação global entre as sequência, calculamos os valores de d_{max} para cada um dos $\binom{6}{2}$ pares de sequências. Nessa etapa, obtivemos que entre as três sequências de Tóquio LC011945, LC11946 e LC11947 obtivemos $d_{max} = 0$.

Portanto, as três sequências de Tóquio serão representadas por LC011945 [14]. Assim, iremos trabalhar apenas com 4 (quatro) sequências (ao invés das 6 iniciais) a saber: LC011945, LC011948, LC011949 e LC016760. A Tabela 25 exibe os valores calculados de d_{max} entre cada um dos $\binom{4}{2}$ pares de sequências, ou seja, para cada um dos $\binom{4}{2}$ pares calculamos a medida d_s, para cada estado $s \in \mathcal{S}$ e, então, consideramos o seu máximo. A memória usada, no nosso caso, foi $o = 3$.

[13] Adiante estão representadas as bases entre as posições 61 e 121. A sequência completa pode ser obtida em <https://www.ncbi.nlm.nih.gov/nuccore/LC011949>.

[14] A escolha foi arbitrária. Poderíamos ter escolhido LC11946 ou LC11947.

Com base na Tabela 25 observa-se que o menor valor de d_{max} ocorre entre as sequências LC011945 e LC011948. Já o maior valor de d_{max} é dado entre a sequência LC011949 em relação às sequências LC011945, LC011948 e LC016760, respectivamente. Com base na matriz apresentada na Tabela 25, podemos construir dendrogramas. Um dendrograma obtido, usando a função de ligação de *Ward* é apresentado na Figura 6. Geramos outros dendrogramas, ver Figura 7, usando funções de ligações diferentes (média, mediana, simples e completa) e, todos eles, exibiram a mesma organização entre as 4 (quatro) sequências, ou seja, 2 (dois) grupos onde LC011949 é um grupo e LC011945, LC011948 e LC016760, outro.

Tabela 25 – Matriz de distâncias, cujas entradas são d_{max} (ver Definição 2) obtidas das sequências LC011945, LC011948, LC011949 e LC016760. Adaptado de (CORDEIRO et al., 2019a) (vide apêndice).

	LC011945	LC011948	LC011949
LC011948	0,00058		
LC011949	0,04204	0,04204	
LC016760	0,00145	0,04204	0,04204

De acordo com dendrograma apresentado na Figura 6, observa-se claramente uma homogeneidade entre 3 (três) das 4 (quatro) sequências: LC011945, LC011948 e LC016760. Além disso, evidencia uma disparidade entre a sequência LC011949 e o grupo formado pelas 3 (três) outras sequências.

A Tabela 25 mostra também que $d_{max} < 1$ para todos os caso, ou seja, $d_s < 1$ em todos os estados $s \in \mathcal{S}$. Assim, as 4 (quatro) sequências sob investigação, de acordo com o teorema 2, foram geradas por uma mesma lei. Entretanto, observa-se certa heterogeneidade entre elas, tendo em vista as magnitudes dos d_{max} calculados. Portanto, podemos tentar identificar quais as sequências são as mais ou menos representativas, dentre as quatro sequências, que é o assunto abordado na seção adiante.

2.4.2.4 Classificação das sequências através de $\delta(i)$

Nesse momento, classificamos cada uma das sequências utilizando o procedimento 1 utilizando a quantidade $\delta(i)$ (ver Definição 3(i)). A relação das sequências em ordem decrescente de representatividade (da mais para a menos representativa) é a seguinte: (i) LC011945 ($\delta(i) = \delta_{(1)} = 0,00145$), LC011948 ($\delta(i) = \delta_{(2)} = 0,00164$), LC016760 ($\delta(i) = \delta_{(3)} = 0,00164$) e LC011949 ($\delta(i) = \delta_{(4)} = 0,04200$).

A sequência que melhor representa o conjunto das 4 (quatro) sequências utilizadas, com base em $\delta(i)$ é a amostra LC011945 proveniente de Tóquio. Já a mais discrepante, ou seja, a que apresentou o maior valor de $\delta(i)$ é LC011949, sendo essa a amostra menos representativa. Como o valor de $\delta(i)$ para a sequência LC011949 é muito alto em relação aos demais, há indicação de que essa sequência tenha uma origem diferente das demais e,

além disso, o paciente 14-181J provavelmente foi infectado por uma cepa distinta da que infectou os outros pacientes japoneses citados na Tabela 24.

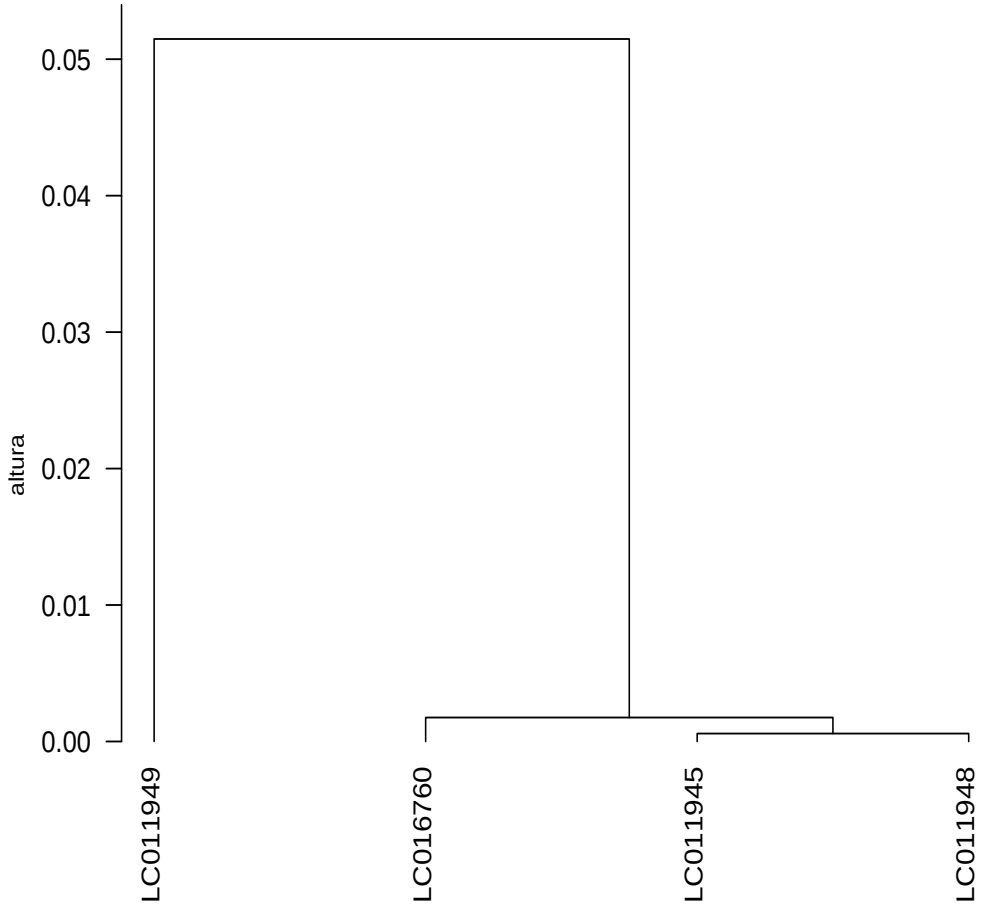


Figura 6 – Dendrograma usando a ligação de *Ward* de uma matriz de distâncias cujas entradas são d_{max} (ver Definição 2), apresentadas na Tabela 25 (CORDEIRO et al., 2019a) (vide apêndice).

A comparação da classificação de LC011949, que tem um $\delta(i)$ de 0,04200, observa-se claramente o impacto de d_{max} nesse valor (ver Tabela 25). Todas as vezes em que se compara a sequência LC011949 com as outras da lista, o valor de d_{max} aumenta pelo menos 25 vezes.

Para identificar de forma mais precisa o significado dessa classificação, comparamos todas as sequências encontradas na base <http://www.ncbi.nlm.nih.gov/> com os perfis de sequências completas do vírus da Dengue tipo 1, do ano de 2014, vindas da Ásia. A relação dos números de acesso dessas sequências é fornecida na Tabela 26. Além disso, para cada uma dessas sequências, incorporamos duas letras ao número, de modo a identificar facilmente o país de origem da sequência sob investigação.

A Figura 3 de (CORDEIRO et al., 2019a) (vide apêndice) apresenta um dendrograma construído tendo por base uma matriz de distância cujas entradas foram os

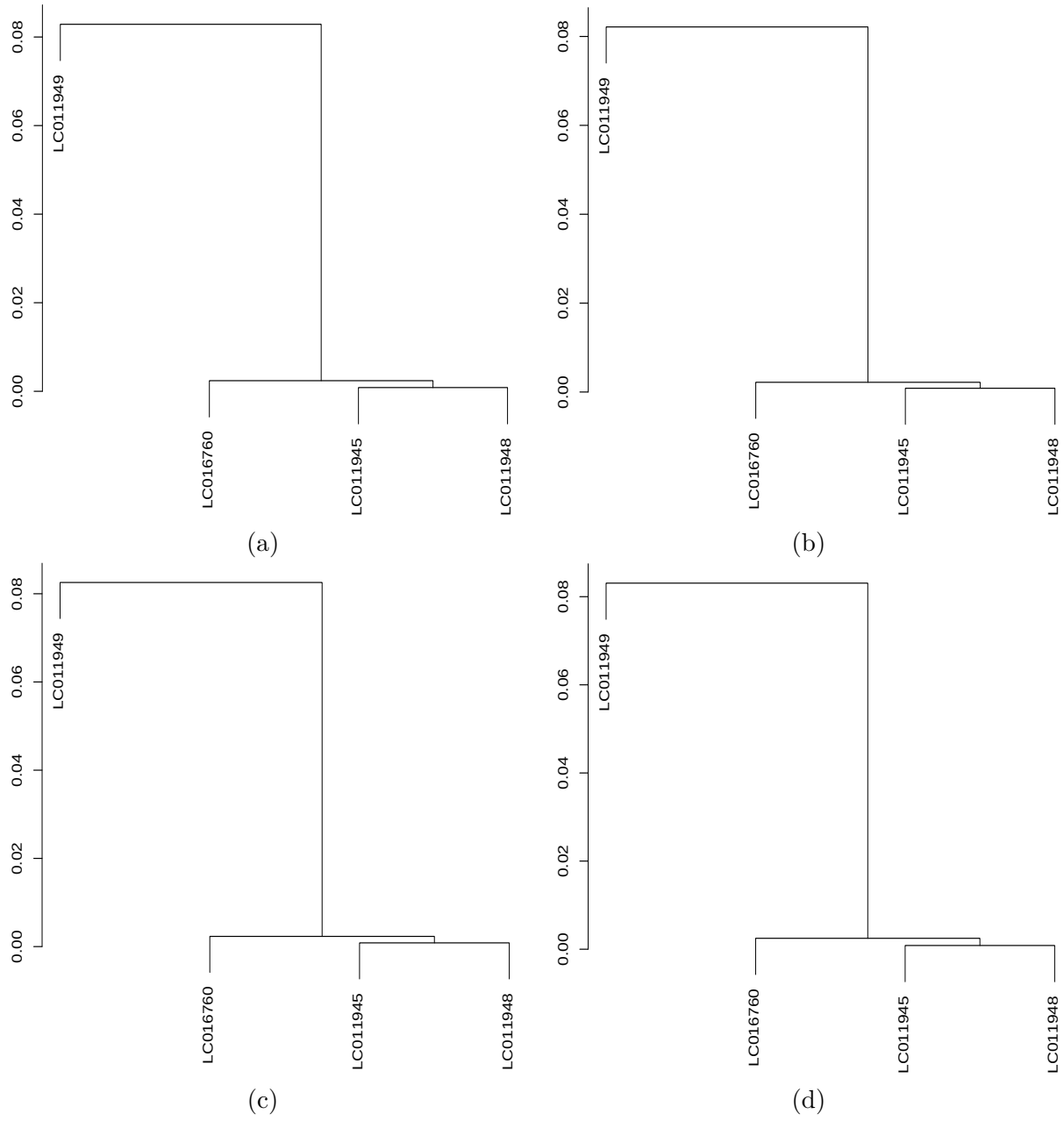


Figura 7 – Dendrogramas usando como matriz de distâncias a da Tabela 25 e as seguintes funções de ligação: (a) média, (b) mediana, (c) simples e (d) completa. Gráficos também disponíveis em <http://www.ime.unicamp.br/~jg/cadvj>.

d_{max} calculados entre todas as sequências listadas na Tabela 26. A função de ligação usada na construção do dendrograma foi a média. O dendrograma circular exibe a sequência japonesa LC011949 com o maior valor de $\delta(i)$ (dentre as sequências LC011945, LC011948, LC011949 e LC016760 utilizadas na análise) em um agrupamento (*cluster*) bastante distante dos outros que compõem o mesmo grupo (ou seja, LC011945, LC011948 e LC016760). Para comentários adicionais sobre o dendrograma ver (CORDEIRO et al., 2019a) (vide apêndice).

Tabela 26 – Relação com os números de acesso (base NCBI) das sequências completas do vírus da Dengue Tipo 1, do ano de 2014, da Ásia. A primeira coluna mostra o país e, a segunda mostra as sequências vindas do país à esquerda. Adaptado de (CORDEIRO et al., 2019a) (vide apêndice).

Origem	Número de acesso da sequência completa				
China (Ch)	KX621252.Ch	KR024708.Ch	MG560268.Ch	KT827379.Ch	KX225489.Ch
	KX225493.Ch	KX459390.Ch	KT827373.Ch	KR024705.Ch	KX459386.Ch
	KX225484.Ch	KT187559.Ch	KX459388.Ch	MG560266.Ch	KX620454.Ch
	KP723476.Ch	MG560269.Ch	KT827377.Ch	KT187563.Ch	MG560267.Ch
	MG560265.Ch	KP723473.Ch	KX225488.Ch	KR024707.Ch	KT827374.Ch
	KX620455.Ch	KX459391.Ch	KX225483.Ch	KX621249.Ch	KT827376.Ch
	KT187560.Ch	KT827371.Ch	KX621250.Ch	KX458014.Ch	KX620451.Ch
	KX459387.Ch	KT187562.Ch	KX458013.Ch	KX620452.Ch	KX225492.Ch
	KX225487.Ch	KT827378.Ch	KX225490.Ch	KT827375.Ch	KX620453.Ch
	KR028435.Ch	KU094071.Ch	KT827372.Ch	KX225491.Ch	KX459389.Ch
	KT187564.Ch	KT187561.Ch	KX459392.Ch	KR024706.Ch	KP686070.Ch
Malásia (Ma)	KU666940.Ma	KX452058.Ma	KU666942.Ma	KX452062.Ma	KX452052.Ma
	KX452053.Ma	KX452065.Ma	KU666941.Ma	KX452056.Ma	KX452063.Ma
	KX452055.Ma	KU666939.Ma	KX452060.Ma	KX452057.Ma	KX452050.Ma
	KX452059.Ma	KX452068.Ma	KX452061.Ma	KX452064.Ma	KX452054.Ma
	KX452051.Ma	KX452067.Ma			
	MF033218.Si	MF033219.Si	MF033216.Si	KX224261.Si	MF033214.Si
	KJ806961.Si	KX224263.Si	MF033217.Si	MF033226.Si	KY921903.Si
	MF033223.Si	MF033225.Si	MF033220.Si	MF033224.Si	MF033215.Si
	MF033221.Si				
Índia (In)	KX618706.In	KT831765.In	KX618705.In		
Japão (Ja)	LC016760.Ja	LC011945.Ja	LC011949.Ja	LC011948.Ja	
Cingapura (Si)	MF033213.Si	KJ806959.Si	KJ806963.Si	MF033227.Si	MF033222.Si
Sri Lanka (Sr)	KP398852.Sr				
Tailândia (Ta)	KU365900.Ta				

2.4.2.5 Conclusão

Nesta seção, após a análise de 6 (seis) cepas de DENV-1 autóctones, que num segundo momento foram restritas a 4 (quatro), identificamos as sequências (cepas) mais e menos representativas deste conjunto. Para esta identificação, utilizamos a quantidade $\delta(i)$ (ver Definição 3 (i)), via Procedimento 1, onde $\delta(i)_{(1)}$ e $\delta(i)_{(n)}$ estão associadas às sequências mais e menos representativas, em termos de lei de formação. A classificação das sequências genômicas do vírus da dengue tipo 1, originados no Japão durante o surto ocorrido em 2014 naquele país foi apresentada, sendo as sequências LC011945 e LC0011949 a mais e menos representativas, respectivamente.

Identificamos as sequências mais representativas do surto, como sendo aquelas vindas de Tóquio (LC011945, LC11946 e LC11947). Além disso, numa comparação com outras cepas, de 2014, observamos uma semelhança entre cepas provenientes da Malásia, Cingapura e China (vide dendrograma constante na Figura 3 de (CORDEIRO et al., 2019a) (vide apêndice)).

De acordo com a classificação que obtivemos e por conta dos grupos obtidos pelos dendrogramas (Figuras 2 e 3 de (CORDEIRO et al., 2019a) (vide apêndice)),

tendemos a concordar com a afirmação de (TAJIMA et al., 2016), no sentido de afirmar que o surto de dengue ocorrido no Japão, em 2014, provavelmente envolveu mais de uma variante do vírus da dengue tipo 1.

Com relação à abordagem que apresentamos nesse trabalho: (i) possibilita quantificar a representatividade das sequências, quando comparadas com um grupo de sequências, (ii) é uma ferramenta de classificação genuinamente estocástica, uma vez que reporta o quão próxima ou distante estão as leis estocásticas de formação das sequências sob consideração.

Como sugestão de pesquisas futuras, poderíamos incluir os demais sorotipos do vírus da Dengue, de modo a (i) estabelecer se a noção d_{max}/d_s é capaz de discriminar entre os demais sorotipos, (ii) identificar o espectro de variações do classificador $\delta(i)$ em cada sorotipo, (iii) investigar o impacto da constante α usada no cálculo de d_s, e consequentemente de d_{max}, em (i) e (ii).

2.5 Conclusão do Capítulo 2

Neste capítulo foi apresentado o processo de construção da métrica d_s, ou seja, mostramos como a métrica foi obtida através do BIC. Na sequência, a métrica d_s e a noção $d_{\max}$ foram definidas e, um exemplo do seu cálculo foi mostrado. Na seção 2.2 mostramos algumas das propriedades teóricas de d_s e $d_{\max}$, a saber: (i) o resultado que garante que d_s é, de fato, uma métrica, (ii) a relação existente entre o BIC e d_s e, (iii) a consistência local e global de d_s e $d_{\max}$, respectivamente.

Já na seção 2.3 exibimos o uso da métrica d_s (e de $d_{\max}$) como um classificador, através das quantidades $\delta(i)$ (Definição 3 (i)) e $\beta(i)$ (Definição 3 (ii)). Um exemplo de cálculo de $\delta(i)$ e $\beta(i)$ (Exemplo 4) e, resultados que sustentam o uso de $\delta(i)$ como um classificador também são apresentados.

Na seção 2.4 foram apresentadas duas aplicações da métrica d_s, em problemas reais, em trabalhos já aceitos para publicação ou publicados. Na subseção 2.4.1, as quantidades $\delta(i)$ e $\beta(i)$ foram empregadas na classificação de sequências de DNA do vírus Zika. O objetivo deste trabalho foi a obtenção de uma "cepa padrão", no que diz respeito a lei geradora destas sequências. Já na subseção 2.4.2, a quantidade $\delta(i)$ foi empregada na classificação de sequências de DNA do vírus da dengue tipo 1 autóctones, que circularam no Japão, durante um surto ocorrido em 2014 naquele país.

Finalmente, como perspectiva de trabalho futuro, temos a intenção de desenvolver um teste de hipóteses utilizando d_s e d_{max}, como estatísticas do teste para testar as hipóteses $H_0 : P^1(\cdot|s) = P^2(\cdot|s)$ versus $H_1 : \exists a \in A : P^1(a|s) \neq P^2(a|s)$, respectivamente. Tal teste seria útil na comparação do comportamento estocástico de processos markovianos

definidos num mesmo alfabeto e, com mesma ordem. As dificuldades envolvidas neste procedimento está, dentre outros, no desconhecimento da distribuição amostral das v.a. d_s e d_{max}. Assim, a viabilização do teste, sem este conhecimento, exigirá o uso de outras metodologias, por exemplo, a reamostragem bootstrap.

3 Modelo ótimo para um conjunto de processos markovianos

No capítulo anterior, tínhamos uma coleção de sequências e, os nossos objetivos eram: (i) identificar a lei de formação "dominante" (a chamada lei *majoritária*, Definição 4), (ii) assumindo a existência da lei majoritária, classificar as sequências quanto a qualidade da representação da lei probabilística destas amostras (qual é a melhor representante, a segunda melhor representante e assim sucessivamente.). Todavia, ao escolhermos uma (ou até mesmo algumas) sequência como uma espécie de "padrão ouro" de um determinado grupo, estamos inevitavelmente perdendo informações. Mesmo que sequências não possuam uma mesma lei de formação, nada impede que elas possuam as mesmas probabilidades de transição, em determinados estados (na verdade, o compartilhamento de probabilidades de transição de alguns estados é até mesmo esperado, uma vez que, em grande parte das vezes, tais sequências são obtidas (geradas) de uma fonte comum. Por exemplo, estas sequências poderiam ser: as sequências de DNA de um determinado vírus, os preços de papeis negociados numa determinada bolsa, etc. O conhecimento nesta redundância, no que diz respeito a probabilidades de transição para alguns estados e alguns processos, poderia ser utilizado na obtenção de um modelo mais parcimonioso, uma vez que não haveria sentido de estimar-se probabilidades, que sabe-se ser iguais, duas ou mais vezes.

Suponhamos novamente, que possuímos uma coleção de sequências, ou seja, uma coleção de amostras $\left\{x_{j,1}^{n_j}\right\}_{j=1}^{p}$ de processos estocásticos $\{(X_{j,t})\}_{j=1}^{p}$. Uma segunda questão que se coloca é a suposição que tais sequências ainda que possuam leis distintas, possuem probabilidades de transição comum para alguns estados de algumas sequências. Com base nisso, queremos obter um modelo único em que, estados e sequências que possuem uma mesma lei, localmente, são considerados conjuntamente. Para esse propósito, a métrica que propusemos em (CORDEIRO et al., 2020) [1] pode ser empregada. Detalharemos algumas das demonstrações dos resultados introduzidos em (CORDEIRO et al., 2020). Apresentamos duas aplicações do uso da métrica: (i) estudo do perfil estocástico de sequências do vírus Epstein-Barr (EBV) em (CORDEIRO et al., 2019b), (ii) caracterização da estrutura genética do vírus Zika (ZV) em (CORDEIRO et al., 2020)[2].

Consideremos $\mathcal{F}$ como sendo uma família de processos Markovianos *relacionados* sob um mesmo alfabeto A, $|A| < \infty$, e uma amostra de tamanho finito de cada um desses processos. Entendemos por relacionados processos que compartilham algumas de suas probabilidades de transição. Em (CORDEIRO et al., 2020) desenvolvemos uma família

[1] Os trabalhos (CORDEIRO et al., 2020) e (CORDEIRO et al., 2019b) também fazem parte da minha contribuição.

[2] Vide nota 1.

para esse tipo de situação. Mostramos também um método de seleção de modelos para a escolha de um modelo conjunto ótimo da família. Por ótimo, compreende-se um modelo mínimo que capture todas as características significantes dos processos em questão. O procedimento é baseado no critério BIC.

Neste capítulo, os seguintes tópicos serão abordados. Na seção 3.1 será definida a noção de equivalência (Definição 5) e o conceito de partição Markoviana (Definição 6). Na seção 3.2, o processo de construção da métrica (Definição 8), tal como um exemplo de como ela é calculada são apresentados. A seção 3.3 trás algumas das propriedades teóricas (como algumas demonstrações) da métrica. A seção 3.4, nas suas subseções, trata de dois estudo de caso, já publicados, contendo aplicações reais da métrica e do modelo de Markov de partições (MMP). Finalmente, a seção 3.5 contém a conclusão do capítulo 3.

3.1 Fundamentação teórica

Seja $\{(X_{j,t})\}_{j=1}^{p}$ uma coleção de p cadeias de Markov, a tempo discreto, assumidas como sendo independentes e definidas em um mesmo alfabeto finito A, ou seja, $|A| < \infty$. Por questões de simplicidade notacional, assume-se que todos os processos têm uma mesma memória o, $o < \infty$. O espaço de estados de cada uma das p cadeia de Markov é dado por $\mathcal{S} = A^o$. Utilizaremos aqui a mesma notação da seção 2.1 do capítulo 2. Para cada $j \in J = \{1, ..., p\}$, $a \in A$ e $s \in \mathcal{S}$, $P^j(s) = P(X_{j,t-o}^{t-1} = s)$ e $P^j(a|s) = P(X_{j,t} = a|X_{j,t-o}^{t-1} = s)$. Além disso, $M = J \times \mathcal{S}$.

Definição 5. *Os elementos $(i, s), (j, r) \in M$ são ditos equivalentes (notação: $(i, s) \sim (j, r)$) se, e somente se, $P^i(a|s) = P^j(a|r)$, para $\forall a \in A$.*

Nota 1. *Para elementos $(i, s), (j, r) \in M$ que não atendam a Definição 5 será usada a seguinte notação: $(i, s) \nsim (j, r)$.*

Exemplo 5. *Considere uma coleção de $p = 4$ cadeias de Markov de ordem $o = 2$, sob um alfabeto $A = \{0, 1, 2\}$ e espaço de estados $\mathcal{S} = \{00, 01, 02, 10, 11, 12, 20, 21, 22\}$ onde as probabilidades de transição $\left\{P^j(a|s), a \in A, s \in \mathcal{S}\right\}_{j=1}^{4}$ são fornecidas na Tabela 27. Nesse caso, o conjunto $M = J \times \mathcal{S}$ é dado por*

$$\begin{aligned} M = J \times \mathcal{S} = \; & \{(1,00), (1,01), (1,02), (1,10), (1,11), (1,12), \\ & (1,20), (1,21), (1,22), (2,00), (2,01), (2,02), \\ & (2,10), (2,11), (2,12), (2,20), (2,21), (2,22), \\ & (3,00), (3,01), (3,02), (3,10), (3,11), (3,12), \\ & (3,20), (3,21), (3,22), (4,00), (4,01), (4,02), \\ & (4,10), (4,11), (4,12), (4,20), (4,21), (4,22)\} \end{aligned}$$

Tabela 27 – Probabilidades de transição de $p = 4$ (quatro) cadeias de Markov de ordem $o = 2$, alfabeto $A = \{0, 1, 2\}$ e espaço de estados $\mathcal{S} = \{00, 01, 02, 10, 11, 12, 20, 21, 22\}$.

s	$P^1(0\|s)$	$P^1(1\|s)$
00	0,20	0,50
01	0,30	0,20
02	0,20	0,50
10	0,30	0,50
11	0,50	0,20
12	0,23	0,47
20	0,50	0,20
21	0,50	0,20
22	0,30	0,20

s	$P^2(0\|s)$	$P^2(1\|s)$
00	0,50	0,30
01	0,20	0,50
02	0,50	0,20
10	0,25	0,35
11	0,30	0,20
12	0,30	0,20
20	0,30	0,20
21	0,30	0,50
22	0,50	0,30

s	$P^3(0\|s)$	$P^3(1\|s)$
00	0,50	0,20
01	0,30	0,50
02	0,30	0,50
10	0,50	0,30
11	0,50	0,20
12	0,20	0,50
20	0,19	0,51
21	0,30	0,50
22	0,20	0,30

s	$P^4(0\|s)$	$P^4(1\|s)$
00	0,18	0,62
01	0,50	0,30
02	0,20	0,50
10	0,30	0,20
11	0,30	0,50
12	0,20	0,30
20	0,30	0,20
21	0,20	0,50
22	0,50	0,30

Tendo por base a Definição 5, podemos observar algumas relações de equivalência entre os elementos do conjunto M. Por exemplo: $(1, 00) \sim (1, 02)$, $(1, 00) \sim (3, 12)$, $(1, 00) \sim (2, 01)$, $(1, 11) \sim (2, 02)$, *entre muitas outras.*

Teorema 7. *A relação de equivalência apresentada na Definição 5 é reflexiva, simétrica e transitiva.*

Prova:

(i) reflexiva: $(i, s) \sim (i, s)$ uma vez que $P^i(a|s) = P^i(a|s)$, para $\forall a \in A$.

(ii) simétrica: Se $(i, s) \sim (j, r)$ então $P^i(a|s) = P^j(a|r)$. Mas, como $P^i(a|s) = P^j(a|r)$ é igual a $P^j(a|r) = P^i(a|s)$, então $(j, r) \sim (i, s)$.

(iii) transitiva: Se $(i, s) \sim (j, r)$ e $(j, r) \sim (k, q)$ então $P^i(a|s) = P^j(a|r)$ e $P^j(a|r) = P^k(a|q)$. Mas, como $P^i(a|s) = P^j(a|r)$ e $P^j(a|r) = P^k(a|q)$, então $P^i(a|s) = P^k(a|q)$. Portanto, $(i, s) \sim (k, q)$.

□

Definição 6. *A coleção $\mathcal{F}$ tem como partição markoviana $\mathcal{L} = \{L_1, ..., L_k\}$ se $\mathcal{L}$ é uma partição de M determinada através da equivalência $\sim$.*

Observação 3. *Com base na Definição 6, podemos pensar que (i, s) e (j, r) estão numa mesma parte caso compartilhem o mesmo mecanismo aleatório de escolha do elemento seguinte na sequência.*

Portanto, o modelo é completamente especificado uma vez que $\mathcal{L}$ tenha sido estimado. Além disso, uma vez que $\mathcal{L}$ já esteja estimado, o conjunto das probabilidades condicionais para a estrutura $\mathcal{L}$ pode ser obtida.

Definição 7. *Se $\mathcal{F}$ possui uma partição markoviana $\mathcal{L} = \{L_1, ..., L_k\}$, então para $\forall L \in \mathcal{L}$, teremos $P_L(a) = P^i(a|s)$, para $\forall (i, s) \in L$ e $\forall a \in A$.*

Exemplo 6. *Considere novamente o Exemplo 5. O conjunto M poderia ser particionado, usando a relação de equivalência apresentada na Definição 5. O conjunto M apresenta a partição markoviana $\mathcal{L} = \{L_1, ..., L_k\}$ onde $k = 10$ e, as partes são dadas na Tabela 28. Algumas de suas partes seriam:*

1. *$(1, 00)$, $(1, 02)$, $(2, 01)$, $(3, 12)$, $(4, 02)$ e $(4, 21)$ compõem uma parte associada a seguinte probabilidade de transição:*

$$P^1(0|00) = P^1(0|02) = P^2(0|01) = P^3(0|12) = P^4(0|02) = P^4(0|21) = 0,20 \quad e$$

$$P^1(1|00) = P^1(1|02) = P^2(1|01) = P^3(1|12) = P^4(1|02) = P^4(1|21) = 0,50.$$

2. *$(1, 10)$, $(2, 21)$, $(3, 01)$, $(3, 02)$, $(3, 21)$ e $(4, 11)$ compõem uma parte associada a seguinte probabilidade de transição:*

$$P^1(0|10) = P^2(0|21) = P^3(0|01) = P^3(0|02) = P^3(0|21) = P^4(0|11) = 0,30 \quad e$$

$$P^1(1|10) = P^2(1|21) = P^3(1|01) = P^3(1|02) = P^3(1|21) = P^4(1|11) = 0,50.$$

Repare que para a modelagem de $p = 4$ cadeias de Markov de ordem $o = 2$ e espaço de estados $\mathcal{S} = \{00, 01, 02, 10, 11, 12, 20, 21, 22\}$ são necessários $|\mathcal{S}|(|A| - 1)p = 72$ parâmetros. No entanto, considerando a partição markoviana teríamos apenas $|\mathcal{L}|(|A|-1) = 20$ parâmetros, ou seja, temos então um modelo mais parcimonioso.

Tabela 28 – Partição Markoviana do conjunto M dos processos apresentados nos Exemplos 5 e 6

$L \in \mathcal{L}$	$P_L(0)$	$P_L(1)$
$\{(1,00),(1,02),(2,01),(3,12),(4,02),(4,21)\}$	0,20	0,50
$\{(1,01),(1,22),(2,11),(2,12),(2,20),(4,10),(4,20)\}$	0,30	0,20
$\{(1,10),(2,21),(3,01),(3,02),(3,21),(4,11)\}$	0,30	0,50
$\{(1,11),(1,20),(1,21),(2,02),(3,00),(3,11)\}$	0,50	0,20
$\{(2,00),(2,22),(3,10),(4,01),(4,22)\}$	0,50	0,30
$\{(3,22),(4,12)\}$	0,20	0,30
$\{(1,12)\}$	0,23	0,47
$\{(2,10)\}$	0,25	0,35
$\{(3,20)\}$	0,19	0,51
$\{(4,00)\}$	0,18	0,62

3.2 Construção de uma métrica

Tendo a família de modelos sido definida, o problema passa a ser o desenvolvimento de um método para a escolha consistente de um modelo com base na amostra disponível. Sejam $\{x_{j,1}^{n_j}\}_{j=1}^{p}$ amostras independentes da coleção de processos $\mathcal{F} = \{(X_{j,t})\}_{j=1}^{p}$ tomando valores no alfabeto finito A, com memória o ($o < \infty$), com tamanhos $\{n_j\}_{j=1}^{p}$ e $n_{\min} = \min\{n_j\}_{j=1}^{p}$.

Em uma dada amostra $x_{i,1}^{n_i}$, proveniente do processo estocástico $(X_{i,t}) \in \mathcal{F}$, $a \in A$ e $s \in \mathcal{S}$,

$$N((i,s),a) := \sum_{z=1}^{n_i-o} \mathbb{1}_{\left\{x_{i,z}^{z+o-1} = s,\ x_{i,z+o} = a\right\}}, \tag{3.1}$$

$$N((i,s)) := \sum_{z=1}^{n_i-o} \mathbb{1}_{\left\{x_{i,z}^{z+o-1} = s\right\}}, \tag{3.2}$$

ou seja, $N((i,s))$ é o total de ocorrências de $s \in \mathcal{S}$ na amostra $x_{i,1}^{n_i}$, $N((i,s),a)$ o total de ocorrências de $s \in \mathcal{S}$ seguidos de $a \in A$ na amostra $x_{i,1}^{n_i}$. [3] Além disso,

$$N(L) = \sum_{(i,s)\in L} N(i,s), L \in \mathcal{L}$$

é o total de ocorrências dos elementos em L e,

$$N(L,a) = \sum_{(i,s)\in L} N((i,s),a)$$

é o total de ocorrências dos elementos em L, seguidos por $a \in A$.

[3] A quantidade $N((i,s))$ considerando a seção 2.1 do capítulo 2 seria escrita como $N_{n_i}(s)$. A opção por $N((i,s))$ ao invés de $N_{n_i}(s)$ se deu pelo fato de que, agora, o espaço em que trabalhamos é o $M = \{1,...,p\} \times \mathcal{S}$ ao invés do $\mathcal{S}$. A omissão de n_i no índice de $N((i,s))$, i.e., $N_{n_i}((i,s))$, foi feita para não carregar a notação.

O nosso método baseia-se no critério BIC. Consideremos as amostras $\{x_{j,1}^{n_j}\}_{j=1}^{p}$ para a coleção de processos $\mathcal{F}$ com partição markoviana $\mathcal{L} = \{L_1, ..., L_k\}$. Como os processos são independentes, a verossimilhança é dada por:

$$P(\{x_{j,1}^{n_j}\}_{j=1}^{p}) = \prod_{j=1}^{p} P^j(x_{j,1}^{n_j}) = \prod_{j=1}^{p} P^j(x_{j,1}^{o}) \prod_{L\in\mathcal{L}} \prod_{a\in A} P_L(a)^{N(L,a)} \quad . \tag{3.3}$$

A pseudo-log-verossimilhança[4] é dada por:

$$\sum_{L\in\mathcal{L}} \sum_{a\in A} N(L,a) \ln(P_L(a)), \tag{3.4}$$

sendo seu máximo atingido em:

$$\sum_{L\in\mathcal{L}} \sum_{a\in A} N(L,a) \ln \frac{N(L,a)}{N(L)} \quad . \tag{3.5}$$

O estimador $\hat{\mathcal{L}}_n$ da partição markoviana $\mathcal{L}$ usando o BIC, associado às amostras para a partição $\mathcal{L}$ de M é dado por

$$\hat{\mathcal{L}}_n = \arg\max_{\mathcal{L}} \mathrm{BIC}(\mathcal{L}, \{x_{j,1}^{n_j}\}_{j=1}^{p}) \quad , \tag{3.6}$$

onde

$$\mathrm{BIC}(\mathcal{L}, \{x_{j,1}^{n_j}\}_{j=1}^{p}) = \sum_{L\in\mathcal{L}} \sum_{a\in A} N(L,a) \ln \frac{N(L,a)}{N(L)} - \frac{(|A|-1)}{2} |\mathcal{L}| \ln \left(\sum_{j=1}^{p} n_j \right). \tag{3.7}$$

A maximização deveria ser feita sob o conjunto de todas as partições de M, conjunto este que, em geral, é muito grande. Ao invés disso, construiremos a partição ótima localmente segundo a seguinte noção de divergência.

Definição 8. *Sejam $\left\{x_{j,1}^{n_j}\right\}_{j=1}^{p}$ amostras de uma coleção $\mathcal{F} = \{(X_{j,t})\}_{j=1}^{p}$ de p cadeias de Markov independentes de tempo discreto, que assumem valores num mesmo alfabeto A*

[4] Entende-se por pseudo-log-verossimilhança a função que é proprorcional a log-verossimihança, ou seja, a log-verossilhança descontadas os logaritmos neperianos (naturais) das probabilidades de ocorrências dos estados iniciais das amostras sob consideração, ou seja, $\sum_{j=1}^{p} \ln P^j(x_{j,1}^{o})$.

com ordem o $(|A|, o < \infty)$*. Para* (i,s)*,* $(j,r) \in M$*, onde* $M = \{1,...,p\} \times A^o$*, definimos*

$$d((i,s),(j,r)) = \\ = \frac{2}{(|A|-1)\ln(\sum_{l=1}^{p} n_l)} \sum_{a \in A} \left\{ N((i,s),a) \ln \frac{N((i,s),a)}{N(i,s)} + N((j,r),a) \ln \frac{N((j,r),a)}{N(j,r)} - \right. \\ \left. N(\{(i,s),(j,r)\},a) \ln \frac{N(\{(i,s),(j,r)\},a)}{N(\{(i,s),(j,r)\})} \right\}, \tag{3.8}$$

onde $N(\{(i,s),(j,r)\}) = N(i,s) + N(j,r)$ *e* $N(\{(i,s),(j,r)\},a) = N((i,s),a) + N((j,r),a)$

Exemplo 7. *Considere* $p = 3$ *amostras provenientes de processos markovianos de ordem* $o = 2$*, definidos num alfabeto* $A = \{0,1\}$ *e cujas probabilidades de transição são determinadas pela Tabela 29. O conjunto* $M = \{1,2,3\} \times \{0,1\}^2$*, possui elementos equivalentes (Definição 5) e, assim, de acordo com a Definição 7 possui uma partição markoviana* $\mathcal{L}$*. As partes* L_k *da verdadeira partição Markoviana e suas respectivas probabilidades condicionais de transição são dadas na Tabela 30.*

Tabela 29 – Probabilidades de transição $P^j(0|s)$ e $P^j(1|s)$ do processo $(X_{j,t})$ onde $j = 1,2,3$.

s	$P^1(0\|s)$	$P^1(1\|s)$
00	0,50	0,50
01	0,30	0,70
10	0,20	0,80
11	0,30	0,70

s	$P^2(0\|s)$	$P^2(1\|s)$
00	0,30	0,70
01	0,50	0,50
10	0,50	0,50
11	0,30	0,70

s	$P^3(0\|s)$	$P^3(1\|s)$
00	0,50	0,50
01	0,30	0,70
10	0,10	0,90
11	0,20	0,80

Três amostras $x_{1,1}^{1001}$*,* $x_{2,1}^{1002}$ *e* $x_{3,1}^{1003}$*, foram geradas à partir de* $P^1(.|s)$*,* $P^2(.|s)$ *e* $P^3(.|s)$*, respectivamente (vide Tabela 29). Estas amostras são apresentadas nas Tabelas B.1, B.2 e B.3.*

Já a Tabela 31 fornece as quantidades $N((i,s),0)$ *e* $N((i,s),1)$ *para a amostra* $x_{i,1}^{n_i}$*, onde* $i = 1,2,3$*. Por exemplo, na amostra* $x_{1,1}^{n_1}$ *o número de vezes que o estado* $s = 10$ *aparece seguido do elemento* $a = 1 \in A$ *é igual a* $N((1,10),1) = 183$ *enquanto que na amostra* $x_{3,1}^{n_3}$ *o número de vezes que o estado* $s = 00$ *aparece seguido do elemento* $a = 1 \in A$ *é igual a* $N((3,00),1) = 16$*. As contagens desta tabela serão utilizadas no cálculo de* $d((i,s),(j,r))$*.*

Tabela 30 – Partição verdadeira do conjunto $M = \{1,2,3\} \times \{0,1\}^2$ onde as probabilidade de transição para os processos $(X_{j,t})$, $j = 1,2,3$ são apresentadas na Tabela 29.

k	$L_k \in \mathcal{L}$	$P(0\|L_k)$	$P(1\|L_k)$
1	(1,00), (2,01), (2,10), (3,00)	0,50	0,50
2	(1,01), (1,11), (2,00), (2,11), (3,01)	0,30	0,70
3	(1,10), (3,11)	0,20	0,80
4	(3,10)	0,10	0,90

Tabela 31 – Quantidades $N((i,s),0)$ e $N((i,s),1)$ para as amostras $x_{1,1}^{n_1}$ (Tabela B.1), $x_{1,1}^{n_1}$ (Tabela B.2) e $x_{1,1}^{n_1}$ (Tabela B.3), sendo n_1, n_2 e n_3 iguais a 1001, 1002 e 1003, respectivamente.

s	$N((1,s),0)$	$N((1,s),1)$
00	39	36
01	65	154
10	35	183
11	153	334

s	$N((2,s),0)$	$N((2,s),1)$
00	53	106
01	95	112
10	106	101
11	112	315

s	$N((3,s),0)$	$N((3,s),1)$
00	16	16
01	52	116
10	16	152
11	116	517

Na Tabela 32 constam as distâncias $d((i,s),(j,r))$ entre os elementos (i,s) e (j,r) do conjunto $M = \{(1,00),(1,01),...,(3,10),(3,11)\}$. Por exemplo, a distância entre os processos considerando o estado $s = 10$ da amostra 1 e o estado $r = 11$ da amostra 3 é igual a $d((1,10),(3,11)) = 0,07275$.

Repare que ao considerarmos elementos (i,s) e (j,r) pertencentes a uma mesma parte, i.e., $(i,s) \sim (j,r)$ obtemos distâncias próximas de 0 (zero), ao passo que as distâncias entre $(i,s) \nsim (j,r)$ serão "mais distantes" de 0 (zero). Por exemplo, os elementos $(1,00)$, $(2,01)$, $(2,10)$ e $(3,00)$ pertencem todos a uma mesma parte (Tabela 30). As distâncias entre esses elementos estão indicadas na Tabela 32 com $\boxdot$ ao lado do valor e, o maior deles é igual a 0,14617 (entre $(2,01)$ e $(2,10)$), i.e., um valor pequeno. A título de exemplo, a distância entre $(1,00)$ e qualquer outro (j,r) que não esteja nesta parte, é muito superior a 0, onde, a menor é entre $(1,00)$ e $(2,00)$ igual a 0,91938.

Tabela 32 – Valores $d((i,s),(j,r))$ calculados entre os elementos (i,s) e (j,r), onde $(i,s),(j,r) \in M$, usando $x_{1,1}^{1001}$, $x_{2,1}^{1002}$ e $x_{3,1}^{1003}$ dadas pelas Tabelas B.1, B.2 e B.3, respectivamente, contidas no anexo. Nota: ⊡, ⊟ e ⊠ indicam distâncias entre elementos de L_1, L_2 e L_3, respectivamente, indicadas na Tabela 30.

	(1,00)	(1,01)	(1,10)	(1,11)	(2,00)	(2,01)
(1,00)	-	-	-	-	-	-
(1,01)	1,47705	-	-	-	-	-
(1,10)	4,39796	1,45304	-	-	-	-
(1,11)	1,45754	⊟ 0,02676	2,41919	-	-	-
(2,00)	0,91938	⊟ 0,07128	1,89825	⊟ 0,02518	-	-
(2,01)	⊡ 0,10270	1,49591	5,71185	1,62675	0,74125	-
(2,10)	⊡ 0,00173	2,58294	7,63857	2,98536	1,47556	⊡ 0,14617
(2,11)	2,33543	⊟ 0,10737	1,10703	⊟ 0,37257	⊟ 0,35356	2,99246
(3,00)	⊡ 0,00448	0,62211	2,05959	0,55477	0,38806	⊡ 0,02341
(3,01)	1,20500	⊟ 0,00910	1,49718	⊟ 0,00157	⊟ 0,02652	1,09522
(3,10)	6,29312	3,12858	0,45326	4,50643	3,61113	8,03448
(3,11)	4,66197	1,48786	⊠ 0,07275	3,20672	1,96753	7,28768

	(2,10)	(2,11)	(3,00)	(3,01)	(3,10)
(1,00)	-	-	-	-	-
(1,01)	-	-	-	-	-
(1,10)	-	-	-	-	-
(1,11)	-	-	-	-	-
(2,00)	-	-	-	-	-
(2,01)	-	-	-	-	-
(2,10)	-	-	-	-	-
(2,11)	4,71468	-	-	-	-
(3,00)	⊡ 0,00202	0,94345	-	-	-
(3,01)	1,97384	⊟ 0,16583	0,51985	-	-
(3,10)	10,06532	2,80936	3,22627	3,11145	-
(3,11)	10,03429	1,16343	1,92368	1,48918	1,03576

3.3 Propriedades teóricas de *d((i,s),(j,r))*

Nesta seção, apresentaremos algumas propriedades de $d((i,s),(j,r))$ tal como suas respectivas demonstrações (ou as etapas destas).

Teorema 8. *Sob as condições da Definição 8,*

1. *a métrica $d((i,s),(j,r))$ é dada em termos da entropia relativa*[5] *por:*

$$d((i,s),(j,r)) = \frac{2}{(|A|-1)\ln(\sum_{l=1}^{p} n_l)} N(i,s) D\left(\frac{N((i,s),\cdot)}{N(i,s)} \,\middle\|\, \frac{N(\{(i,s),(j,r)\},\cdot)}{N(\{(i,s),(j,r)\})}\right) + \tag{3.9}$$

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

2. *Se $P^i(\cdot|s) = P^j(\cdot|r)$. Então:*

   (i)

$$D\left(\frac{N((i,s),\cdot)}{N(i,s)} \,\middle\|\, \frac{N(\{(i,s),(j,r)\},\cdot)}{N(\{(i,s),(j,r)\})}\right) \leqslant D\left(\frac{N((i,s),\cdot)}{N(i,s)} \,\middle\|\, P^i(\cdot|s)\right), \tag{3.10}$$

   (ii)

$$D\left(\frac{N((j,r),\cdot)}{N(j,r)} \,\middle\|\, \frac{N(\{(i,s),(j,r)\},\cdot)}{N(\{(i,s),(j,r)\})}\right) \leqslant D\left(\frac{N((j,r),\cdot)}{N(j,r)} \,\middle\|\, P^j(\cdot|r)\right). \tag{3.11}$$

Prova do Teorema 8:

(i)

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

[5] A entropia relativa entre P e Q, sobre o conjunto A é dada por $D(P(\cdot)\,\|\,Q(\cdot)) = \sum_{a\in A} P(a)\ln\frac{P(a)}{Q(a)}$.

$$= N(i,s)\sum_{a\in A}\frac{N((i,s),a)}{N(i,s)}\ln\frac{\frac{N((i,s),a)}{N(i,s)}}{\frac{N(\{(i,s),(j,r)\},a)}{N(\{(i,s),(j,r)\})}} + N(j,r)\sum_{a\in A}\frac{N((j,r),a)}{N(j,r)}\ln\frac{\frac{N((j,r),a)}{N(j,r)}}{\frac{N(\{(i,s),(j,r)\},a)}{N(\{(i,s),(j,r)\})}}$$

$$= N(i,s)D\left(\frac{N((i,s),\cdot)}{N(i,s)}\,\middle\|\,\frac{N(\{(i,s),(j,r)\},\cdot)}{N(\{(i,s),(j,r)\})}\right)+$$

$$N(j,r)D\left(\frac{N((j,r),\cdot)}{N(j,r)}\,\middle\|\,\frac{N(\{(i,s),(j,r)\},\cdot)}{N(\{(i,s),(j,r)\})}\right)$$

onde $D(P(\cdot)\,\|\,Q(\cdot)) = \sum_{a\in A}P(a)\ln\frac{P(a)}{Q(a)}$, i.e., a entropia relativa entre P e Q, sobre o conjunto A.

(ii) A demonstração se encontra em (CORDEIRO et al., 2020) e segue da seguinte nota:

Nota

Sob a suposição de que $P^i(\cdot|s) = P^j(\cdot|r)$ o estimador de máxima verossimilhança dessa probabilidade é igual a $\frac{N(\{(i,s),(j,r)\},\cdot)}{N(\{(i,s),(j,r)\})}$. Portanto são válidas:

$$\prod_{a\in A}\left(\frac{N(\{(i,s),(j,r)\},a)}{N(\{(i,s),(j,r)\})}\right)^{N((i,s),a)} \geqslant \prod_{a\in A}P^i(a|s)^{N((i,s),a)} \tag{3.12}$$

$$\prod_{a\in A}\left(\frac{N(\{(i,s),(j,r)\},a)}{N(\{(i,s),(j,r)\})}\right)^{N((j,r),a)} \geqslant \prod_{a\in A}P^j(a|r)^{N((j,r),a)} \tag{3.13}$$

□

O critério BIC indica que $(i,s),(j,m) \in M$ deveriam estar numa mesma parte se, e somente se, $d((i,s),(j,r)) < 1$. Portanto, a divergência $d((i,s),(j,r))$ é relacionada ao critério BIC. Essa afirmação é dada por meio do seguinte teorema.

Teorema 9. *Sob as suposições da Definição 8, definamos duas partições de M a saber: (i) $\mathcal{L}^1 = \{L^1_1,...,L^1_k\}$ onde $L^1_1 = \{(i,s),(j,r)\}$, (ii) $\mathcal{L}^2 = \{L^2_1,...,L^2_{k+1}\}$ onde $L^2_1 = \{(i,s)\}$, $L^2_2 = \{(j,r)\}$ e $L^2_m = L^1_{m-1}$, $m = 3,...,k+1$. Portanto, em $\mathcal{L}^1$ temos que $(i,s)\sim(j,r)$ ao passo que, em $\mathcal{L}^2$, $(i,s)\nsim(j,r)$. Além disso, sejam $BIC(\mathcal{L}^1,\{x^{n_j}_{j,1}\}^p_{j=1})$ e $BIC(\mathcal{L}^2,\{x^{n_j}_{j,1}\}^p_{j=1})$ os valores do BIC para os modelos de partições $\mathcal{L}^1$ e $\mathcal{L}^2$, respectivamente. Então:*

$$BIC(\mathcal{L}^1,\{x^{n_j}_{j,1}\}^p_{j=1}) > BIC(\mathcal{L}^2,\{x^{n_j}_{j,1}\}^p_{j=1}) \text{ se, e somente se } d((i,s),(j,r)) < 1$$

Prova do Teorema 9: A ideia da construção de $BIC(\mathcal{L}^1,\{x^{n_j}_{j,1}\}^p_{j=1})$ e $BIC(\mathcal{L}^2,\{x^{n_j}_{j,1}\}^p_{j=1})$ é dada na Seção 3.2. O modelo ótimo é aquele em que se maximiza o BIC. Assim sendo,

se $(i,s) \sim (j,r)$ então o BIC do modelo de partições $\mathcal{L}^1$ deve ser maior que o de $\mathcal{L}^2$. Para mais detalhes vide (CORDEIRO et al., 2020). □

O Teorema 10 prova que $d((i,s),(j,r))$ é estatisticamente consistente, uma vez que, para um n_{min} suficientemente grande, $d((i,s),(j,r))$ é capaz de detectar se as leis sendo comparadas são semelhantes ou discrepantes. Em outras palavras, para um tamanho mínimo de amostra suficientemente grande, $d((i,s),(j,r))$ é capaz de dizer se $(i,s) \sim (j,r)$ ou se $(i,s) \nsim (j,r)$.

Teorema 10. *Sob as suposições da Definição 8, $d((i,s),(j,r))$ é estatisticamente consistente, ou seja, se $n_{\min} = min\{n_1,...,n_p\}$,*

(i) $\lim_{n_{\min}\to\infty} d((i,s),(j,r)) = 0$, *se e somente se,* $P^i(.|s) = P^j(.|r)$,

(ii) $\lim_{n_{\min}\to\infty} d((i,s),(j,r)) = \infty$, *se e somente se,* $P^i(.|s) \neq P^j(.|r)$.

Prova do Teorema 10: Escreve-se $d((i,s),(j,r))$ em termos da entropia relativa (vide (3.9) na página 77). $[(i) \Leftarrow]$ Em $d((i,s),(j,r))$ na forma da entropia relativa, usa-se (3.10) e (3.11), o lema 6.3 de (CSISZÁR; TALATA, 2006) e o lema 6.2, também de (CSISZÁR; TALATA, 2006). $[(ii) \Leftarrow]$ Se $P^i(a|s) = P^j(a|r)$, pode ser mostrado que $D\Big(\frac{N(\star,\cdot)}{N(\star)}\Big\|\frac{N(\{(i,s),(j,r)\},\cdot)}{N(\{(i,s),(j,r)\})}\Big) > 0$. Quando $\min\{n_1,\cdots,n_p\} \to \infty$, mostra-se que $\frac{N(\star)}{\ln(\sum_{l=1}^p n_l)} \to \infty$, onde $\star = (i,s),(j,r)$. $[(i) \Rightarrow]$ Assume-se que existe um $a \in A$ tal que $P^i(a|s) \neq P^j(a|r)$ e, usando $[(ii) \Leftarrow]$ chega-se numa contradição. $[(ii) \Rightarrow]$ assume-se que $P^i(a|s) = P^j(a|r)$ e, por $[(i) \Leftarrow]$ também temos uma contradição. Para mais detalhes da demonstração, vide Teorema 1 de (CORDEIRO et al., 2020). □

Observação 4. *O Teorema 10 (i) pode ser visto como* $\lim_{n_{\min}\to\infty} d((i,s),(j,r)) = 0$, *se e somente se,* $(i,s) \sim (j,r)$. *Já 10 (ii) como* $\lim_{n_{\min}\to\infty} d((i,s),(j,r)) = \infty$, *se e somente se,* $(i,s) \nsim (j,r)$.

Neste trabalho, a todo momento chamamos $d((i,s),(j,r))$ de métrica. No resultado adiante apresentamos o resultado que prova que, de fato, $d((i,s),(j,r))$ é uma métrica definida em M, ou seja, é não-negativa, simétrica e obedece a desigualdade triangular. Assim, M é um espaço métrico.

Teorema 11. *Sob as suposições da Definição 8, d é uma métrica, isto é,*

(i) $d((i,s),(j,r)) \geqslant 0$, *com igualdade quando* $\forall a \in A$, $\frac{N((i,s),a)}{N((i,s))} = \frac{N((j,r),a)}{N((j,r))}$ *(não-negativa);*

(i) $d((i,s),(j,r)) = d((j,r),(i,s))$ *(simetria);*

(i) Dados $(i,s),(j,r),(l,x) \in M$, $d((i,s),(l,x)) \leqslant d((i,s),(j,r)) + d((j,r),(l,x))$ *(desigualdade triangular).*

Prova: [(i)] Escreve-se $d((i,s),(j,r))$ em termos da entropia relativa (vide (3.9) na página 77). De imediato, constata-se que $d((i,s),(j,r)) > 0$. Como $\frac{N(\{(i,s),(j,r)\},a)}{N(\{(i,s),(j,r)\})} = w\frac{N((i,s),a)}{N((i,s))} + (1-w)\frac{N((j,r),a)}{N((j,r))}$, se $\frac{N((i,s),a)}{N((i,s))} = \frac{N((j,r),a)}{N((j,r))}$ então as entropias relativas em (3.9) zeram e $d((i,s),(j,r)) = 0$. [(ii)] segue imediatamente da definição.

[(iii)] Escreve-se a desigualdade

$$d\Big((i,s),(l,x)\Big) \leqslant d\Big((i,s),(j,r)\Big) + d\Big((j,r),(l,x)\Big)$$

como

$$d\Big((i,s),(j,r)\Big) + d\Big((j,r),(l,x)\Big) - d\Big((i,s),(l,x)\Big) \geqslant 0 \tag{3.14}$$

ou seja, a demonstração consiste em mostrar que o lado esquerdo de (3.14) é não-negativo. Isso pode ser feito abrindo-se as expressões $d\Big((i,s),(j,r)\Big)$, $d\Big((j,r),(l,x)\Big)$ e $d\Big((i,s),(l,x)\Big)$. Após algum algebrismo, usando (3.9) (quanto necessário) e, sabendo que

$$\frac{N((l,x),a)N(\{(i,s),(l,x)\})}{N(\{(i,s),(l,x)\},a)} \geqslant \frac{1}{n_i+n_l} \quad \text{e} \quad \frac{N((i,s),a)N(\{(i,s),(l,x)\})}{N(\{(i,s),(l,x)\},a)} \geqslant \frac{1}{n_i+n_l}$$

consegue-se mostrar que o lado esquerdo de (3.14) é, de fato, não-negativo. Vide demonstração do Teorema 2 de (CORDEIRO et al., 2020). □

Até este ponto, pelo Teorema 10 temos que $d((i,s),(j,r))$ é estatisticamente consistente, ou seja, para um n_{min} suficientemente grande, $d((i,s),(j,r))$ é capaz de dizer se $(i,s) \sim (j,r)$ ou se $(i,s) \nsim (j,r)$. Adiante, mostraremos que a partição verdadeira $\mathcal{L}$, obtida da Definição 6, pode ser estimada eventualmente quase certamente, por meio do BIC, ou seja, para um n_{min} suficientemente grande, a partição verdadeira $\mathcal{L}$ pode ser recuperada pelo BIC.

Teorema 12. *Sejam* $\left\{x_1^{l,n_l}\right\}_{l=1}^{p}$ *amostras da coleção* $\mathcal{F}$ *de* p *cadeias de Markov independentes de tempo discreto sobre um mesmo alfabeto* A *e ordem* o *(*$|A|, o < \infty$*). Seja o conjunto* $M = \{1,...,p\} \times A^o$ *e considere* $\widehat{\mathcal{L}} = \arg\max_{\mathcal{L}} BIC(\mathcal{L}, \left\{x_1^{l,n_l}\right\}_{l=1}^{p})$. *Se* $\mathcal{L}^\star$ *é a partição de* M *obtida pela Definição 6, então quando* $n_{\min} = min\{n_1,...,n_p\} \to \infty$, *eventualmente, quase certamente,* $\widehat{\mathcal{L}} = \mathcal{L}^\star$.

Prova: Começa-se identificando o espaço das partições de M onde o BIC máximo ocorre. Pode ser mostrado que o BIC máximo ocorre num espaço de partições, na qual todas as partições possuam suas partes (elementos da partição) contendo elementos de M equivalentes entre si. Em uma partição genérica deste espaço de partições, aplica-se o

Corolário 1 de (GARCÍA; GONZÁLEZ-LÓPEZ, 2017) de modo a reduzir o seu cardinal. O final da demonstração é feito por redução ao absurdo (contradição). Para mais detalhes, vide Teorema 3 de (CORDEIRO et al., 2020). □

3.4 Estudos de caso

3.4.1 Perfil estocástico do vírus Epstein-Barr em tipos de carcinomas nasofaringeais

Nesta seção, descrevo o trabalho (CORDEIRO et al., 2019b) (vide apêndice) já publicado, que integra a minha contribuição para o tema. Nesse estudo, construímos um perfil do vírus Epstein-Barr (EBV) por meio de sequências genômicas obtidas de pacientes portadores de carcinoma nasofaringeal (NPC). Consideramos um conjunto de sequências provenientes da fonte gratuita NCBI. O conjunto de sequências é uma coleção de amostras independentes de processos estocásticos relacionados por uma relação de equivalência dada pela Definição 5.

A proposta principal desse trabalho é a criação de um modelo parcimonioso que permita representar a organização genômica do vírus Epstein-Barr (EBV), considerando sequências de DNA deste, provenientes de pacientes portadores de NPC, doença esta com alta incidência no sul da China. A literatura especializada sugere que o EBV desempenha um papel importante na ocorrência de NPC. Além disso, mesmo o NCP ocorrendo em todo o mundo, apresenta um padrão de ocorrência geográfico bastante evidente (ver (CORDEIRO et al., 2019b)). Nesse trabalho, utilizamos duas sequências completas conhecidas na literatura, a saber: GD1 e GD2 e uma sequência incompleta chamada HKNPC1.

Nesse estudo, consideramos cada uma das sequências como sendo amostras de processos Markovianos. Além disso, para a obtenção de um modelo mais parcimonioso, i.e., com um número menor de parâmetros a serem estimados, utilizamos aqui o modelo de Markov de Partições (MMP) introduzido pela Definição 7.

3.4.1.1 Dados e resultados

3.4.1.1.1 Genoma do EBV

O conjunto de dados utilizado no presente estudo foi obtido no repositório NCBI. A sequência GD1 foi isolada da saliva de um paciente portador de NPC e, as sequências GD2 e HKNPC1, de biópsias específicas realizadas nos pacientes. Para detalhes, ver Tabela 2 em (CORDEIRO et al., 2019b) (vide apêndice).

A representação das sequências é dada por uma sequência das seguintes letras: a (adenina), c (citosina), g (guanina) e t (timina). Cada uma das sequências é vista como

um processo Markoviano, de ordem o ($o < \infty$), assumindo valores num alfabeto A, dado por $A = \{a, c, g, t\}$. Além disso, o conjunto J é dado por $J = \{1, 2, 3\}$, com 1, 2 e 3 referindo-se às sequências GD1, GD2 e HKNPC1, respectivamente. A memória o permitida é tal que $o < log_{|A|}(n) - 1$, sendo $n = \sum_{j \in J} n_j$. No nosso caso, $n = 508695$, o que resulta em $log_{|A|}(n) - 1 = log_4(508695) - 1 = 8,478$, ou seja, $o < 8$. Como na modelagem de problemas de dados genômicos, os elementos de A costumam ser dados em triplas, então ordens o múltiplas de 3 são recomendadas. Este trabalho foi desenvolvido sob modelos de ordem $o = 3$, por questões de simplicidade de exposição ao leitor.

3.4.1.1.2 Estimação

Os resultados obtidos com ordem $o = 3$ são mostrados nas Tabelas 3 e 4 de (CORDEIRO et al., 2019b) (vide apêndice). A Tabela 3 de (CORDEIRO et al., 2019b) apresenta a estimativa $\hat{\mathcal{L}}$ de $\mathcal{L}$, que é a partição markoviana do conjunto

$$M = \{1, 2, 3\} \times \{a, c, g, t\}^3,$$

onde $|\hat{\mathcal{L}}| = 34$. Para cada uma das partes $\hat{L}_i$ estimadas, a sua composição também é informada, listando os elementos da esquerda para a direita, e de cima para baixo, de acordo com a magnitude de d como cada um dos elementos de M compõem uma parte. Na última coluna da Tabela 3 de (CORDEIRO et al., 2019b), o maior valor de d obtido (o maior valor de d obtido na construção da parte) após a aplicação do critério aglomerativo (ver C). Por fim, na Tabela 4 de (CORDEIRO et al., 2019b) são apresentadas as estimativas das probabilidades condicionais para cada uma das partes, onde

$$\hat{P}(\cdot|\hat{L}_i) = \frac{N(\hat{L}_i, \cdot)}{N(\hat{L}_i)} = \frac{\sum_{(j,s)\in\hat{L}_i} N((j,s),\cdot)}{\sum_{(j,s)\in\hat{L}_i} N((j,s))}; \quad \cdot \in \{a, c, g, t\}. \tag{3.15}$$

Ainda com base na Tabela 3 de (CORDEIRO et al., 2019b), observa-se que algumas partes apresentam maior homogeneidade entre seus elementos, uma vez que seu valor máximo de d é muito próximo de zero (esse é o caso, por exemplo, das partes 30, 31, 32, 33 e 34 indicadas na Tabela 3 de (CORDEIRO et al., 2019b)). Algumas outras partes possuem um valor máximo de d muito próximo de 1, que é o máximo permitido para elementos pertencerem a uma mesma parte, o que se traduz em grande diversidade (esse é o caso, por exemplo, das partes 1, 2 e 3 mostradas na Tabela 3 de (CORDEIRO et al., 2019b)).

As probabilidades condicionais de cada uma das partes obtidas também foram estimadas e, seus valores são apresentados na Tabela 4 de (CORDEIRO et al., 2019b). Por exemplo, a parte

$\hat{L}_{27} = \{(1, cgc), (1, ctc), (2, ctc), (3, ctc)\}$ (Tabela 3 de (CORDEIRO et al., 2019b))

possui as seguintes probabilidades condicionais de transição para cada um dos elementos do alfabeto $A = \{a, c, g, t\}$: $P(a|(j,s)) = 0,17706$, $P(c|(j,s)) = 0,40208$, $P(g|(j,s)) = 0,18497$ e $P(t|(j,s)) = 0,23589$, onde $(j,s) \in \hat{L}_{27}$ (vide Tabela 4 de (CORDEIRO et al., 2019b)). Assim, dado que estamos na j-ésima amostra, no estado s e $(j,s) \in \hat{L}_{27}$, então o elemento mais provável de ser o seguinte da realização é o c ($P(c|(j,s)) = 0,40208$) e, o menos provável, o a ($P(a|(j,s)) = 0,17706$).

Observa-se que existe uma tendência em todas as partes, com exceção da parte 9, a escolherem (visitarem) o elemento c (em 14 das partes) ou g (em 19 das partes).

Nesse trabalho também comparamos os aspectos gerais de modelos ajustados, sob três cenários, no que diz respeito à ordem: (i) $o = 3$, (ii) $o = 4$, (iii) $o = 6$ (Tabela 5 de (CORDEIRO et al., 2019b) (vide apêndice)). Observamos que, conforme o esperado, o valor do BIC aumenta à medida que aumenta-se a ordem o do modelo, o que indica um melhor modelo para ordem maiores (neste caso, $o = 6$). No entanto, o modelo de ordem $o = 6$ é muito menos parcimonioso que o modelo de ordem $o = 3$ (haja visto que o modelo de ordem 6 apresenta aproximadamente 4 vezes mais parâmetros que o de ordem 3). Por isso, nesse estudo, optou-se por um modelo de $o = 3$. Os resultados dos modelos possuindo ordens $o = 3, 4$ estão disponíveis em <http://www.ime.unicamp.br/~jg/spebv/>.

3.4.1.1.3 Conclusão

Neste estudo, obtivemos uma representação única das sequências GD1, GD2 e da HKNPC1, por meio de um modelo de Markov de Partições (Definição 7). O modelo obtido contém uma diminuição expressiva na quantidade de parâmetros a serem estimados (de 576 para 102) e, para a estimação destes parâmetros, utilizamos a noção apresentada em (3.15).

Tiramos proveito da similaridade entre as sequências GD1, GD2 e HKNPC, apontada por outros trabalhos (KWOK et al., 2012), para estimar com alta qualidade os parâmetros, de um modelo que englobasse as três sequências num único modelo.

Além disso, obtivemos uma representação da dinâmica deste processo comum, por meio da partição estimada do espaço de estados. Identificamos 34 partes que representam, de forma bastante mais parcimoniosa, a lei de geração das sequências GD1, GD2 e HKNPC e, estas definem o modelo de partição. Ao dispormos das partes, conseguimos distinguir quais estados (associados a suas respectivas amostras) são estocasticamente equivalentes, uma vez que possuem a mesma probabilidade de transição.

Outro achado interessante é o fato de que algumas partes são compostas por elementos mais homogêneos que outras. Isso pode ser observado com base na maior

distância d (Definição 8) entre seus elementos. Assim sendo, o conhecimento de um valor de corte para d é bastante útil (esse valor, pelo Teorema 9 existe e é igual a 1). O modelo único que foi obtido possui muitas das suas partes formadas por estados análogos, entre as diferentes sequências.

3.4.2 Caracterização da estrutura genética do Zika vírus (ZV)

Nesta seção, a aplicação realizada em (CORDEIRO et al., 2020) é descrita, que é parte da minha contribuição para o tema. Em (CORDEIRO et al., 2020), construímos um perfil genético do ZV vindas de regiões tropicais e subtropicais. Tendo por base 44 sequências brasileiras completas de ZV e, num segundo momento, outras 153 sequências de ZV também completas (onde as 44 brasileiras estão incluídas) apresentamos uma representação global destas. Este modelo único levou em consideração às probabilidades de transição redundantes existentes no conjunto de sequências. Este perfil genético de ZV obtido (primeiro para as sequências brasileiras e, depois para as dos demais países) poderia ser usado em comparações futuras.

O ZV, como qualquer outro vírus, possui variações genéticas que, no limite, podem resultar em cepas mutantes. Tais cepas podem (ser ou tornar-se) resistentes a terapias médicas usadas no seu combate. Uma variação genética poderia ser detectada, por exemplo, por meio da comparação de uma sequência de DNA (vinda de um ZV supostamente mutante) com um perfil genético obtido previamente. Assim, a existência deste perfil genético de ZV (e de qualquer outro vírus) seria útil no estudo da evolução genética do vírus em questão. Devemos ter em mente, que uma única sequência de DNA é o resultado de várias interações. Assim, o mecanismo probabilístico de formação desta sequência sofre influência de fatores ambientais e do hospedeiro do qual a sequência foi originada. Ao olharmos para um conjunto de sequências, podemos identificar redundâncias, no que diz respeito às suas leis de formação. Assim, pode-se determinar que tipos de características (aqui, probabilidades de transição de um estado s para um elemento qualquer $a \in A$ do alfabeto) são comuns a todos os indivíduos. Esta identificação é o que entendemos por perfil genético.

As 44 sequências de VZ brasileiras, como as 153 outras (que incluem as 44 países) vindas de outros países americanos são as mesmas utilizadas em (GARCÍA et al., 2018). No entanto, a perspectiva de análise aqui é bastante distinta. Em (GARCÍA et al., 2018), o nosso objetivo era a identificação de uma sequência que melhor representasse o conjunto de todas as sequências onde, esta sequência, seria uma espécie de "padrão-ouro" do conjunto de sequências. Já em (CORDEIRO et al., 2020) não há o descarte de nenhuma das sequências, uma vez que um modelo único que representa, de maneira parcimoniosa, todo o conjunto de sequências é ajustado.

3.4.2.1 Dados e resultados

Os dados consistem dos mesmos usados em (GARCÍA et al., 2018). As 44 sequências genômicas completas provenientes do Brasil, consideradas neste estudo, estão listadas na Tabela B.4. Já as Tabelas B.9 e B.10 apresentam o conjunto das 153 sequências completas vindas de vários países (inclusive as 44 sequências brasileiras). Novamente, cada sequência de DNA foi assumida como sendo a concatenação de elementos do alfabeto $A = \{a, c, g, t\}$. A lei probabilística de formação destas sequências foi assumida como sendo um processo markoviano de ordem (memória) $o = 3$. Para mais detalhes, consulte a seção 5 de (CORDEIRO et al., 2020).

3.4.2.1.1 Sequências brasileiras:

Inicialmente ajustamos o modelo de Markov de Partições (MMP) para as 44 sequências brasileiras. Aplicamos a métrica $d((i, s), (j, r))$, através do método descrito na Seção C, para todos os elementos (i, s), (j, r) em $M = \{1, ..., p\} \times \mathcal{S} = \{1, ..., 44\} \times \{a, c, g, t\}^3$. O conjunto $\{1, ..., 44\}$ faz referência as 44 sequências brasileiras de VZ, de acordo com a Tabela (B.4). Por exemplo, sequência brasileira número 1 = KX197192.1, sequência brasileira número 2 = KY014296.2, sequência brasileira número 3 = KY014297.2, ..., sequência brasileira número 44 = KY817930.1. O ajuste do modelo de partições forneceu um conjunto $\hat{\mathcal{L}}$ com $|\hat{\mathcal{L}}| = 38$. Adiante, tendo as partes L_k do conjunto de partição $\hat{\mathcal{L}}$ de M sido determinadas, calculamos $N(L_k)$ e $N(L_k, x)$ onde $k = 1, 2, ..., 38$ e $x =$ a, c, g, t (Tabela 33). Por exemplo, para $k = 16$, temos que $N(L_{16}) = 10081$, com $N(L_{16}, a)$, $N(L_{16}, c)$, $N(L_{16}, g)$ e $N(L_{16}, a)$ sendo 2934, 2052, 3387 e 1708, respectivamente. O conhecimento de $N(L_k)$ e $N(L_k, x)$ nos permite calcular $\hat{P}(x|L_k) = \dfrac{N(L_k, x)}{N(L_k)}$, $x =$ a, c, g, t (vide Tabela 3 de (CORDEIRO et al., 2020)).

Com base nas distribuições de probabilidade $\hat{P}(x|L_k)$, $x =$ a, c, g, t, apresentadas na Tabela 3 de (CORDEIRO et al., 2020), observamos que a maioria das partes tem a tendência de escolher os elementos a ou g como sendo o próximo elemento da cadeia.

O conjunto M possui um total de $|M| = 44 \times 4^3 = 2816$ elementos o que resulta em $3 \times 2816 = 8448$ parâmetros a serem estimados (que são as probabilidades de transição de cada elemento de M para algum dos elementos do alfabeto A. No entanto, com o modelo de partições temos $|\hat{\mathcal{L}}| = 38$ o que resulta em $3 \times 38 = 114$ parâmetros a serem estimados, ou seja, em torno de 1,4% do número de parâmetros a serem estimados sem a estrutura de partições. Assim, o modelo de partições é bastante parcimonioso.

As Tabelas 34 e 35 fornecem uma descrição das partes obtidas, após o ajuste do modelo de partições, as 44 sequências brasileiras. Por exemplo, na Tabela 34 temos que a parte L_7 é composta por 41 elementos formados e, apenas 4 (quatro) estados (dentre os 64 disponíveis) fazem parte da composição dos elementos de L_7. Além disso, o estado cgg

Tabela 33 – Valores de $N(L_i)$ e $N(L_i, x)$ onde $x = a, c, g, t$, obtidos após o ajuste do MMP para as 44 sequencias brasileiras.

L_i	$N(L_i)$	$N(L_i, a)$	$N(L_i, c)$	$N(L_i, g)$	$N(L_i, t)$
1	10646	3638	1933	3855	1220
2	18688	7480	4424	2794	3990
3	11808	3508	2387	2411	3502
4	5216	2274	921	843	1178
5	23552	9007	5342	3084	6119
6	20206	7559	4059	5727	2861
7	3491	1416	796	782	497
8	25871	5768	4928	11367	3808
9	9208	2528	2173	2326	2181
10	9108	2756	2262	1885	2205
11	13266	3055	3269	3209	3733
12	10724	1996	2766	3194	2768
13	7709	1740	1953	2346	1670
14	8860	2410	2112	2725	1613
15	25397	9572	5110	5992	4723
16	10081	2934	2052	3387	1708
17	13153	1590	2841	4914	3808
18	11227	1564	2458	4601	2604
19	20472	5401	3108	7436	4527
20	12563	3995	3412	1779	3377
21	12230	4322	1945	3363	2600
22	12030	1808	2062	5265	2895
23	4876	430	956	2300	1190
24	8618	2366	2588	2118	1546
25	5170	2245	1025	306	1594
26	8960	1748	2005	3078	2129
27	8824	2956	2294	931	2643
28	2882	425	1037	712	708
29	10527	1760	3158	4015	1594
30	13874	1626	3692	6192	2364
31	8159	1241	1721	3811	1386
32	3241	555	394	1708	584
33	8245	2345	2289	902	2709
34	10847	3081	1924	4292	1550
35	14242	4714	2438	4743	2347
36	2707	546	943	536	682
37	6299	1495	1762	703	2339
38	8770	2084	1625	3478	1583

está presente em 38 dos 41 elementos (92,7%) de L_7, ou seja, 92,7% dos elementos de L_7 contém o estado *cgg* em sua composição. Por fim, temos que 40 (dentre as 44 sequências brasileiras) contribuem com elementos para a partição L_7. A Tabela 35 fornece informação complementar a exibida pela Tabela 34, sendo a única diferença entre as duas é que a Tabela 35 descreve as partições de L_{20} até L_{38}, enquanto que Tabela 34, de L_1 até L_{19}.

Tabela 34 – 1 de 2 - Análise descritiva das partes do modelo de partições estimado para as 44 sequências brasileiras de vírus da Zika. Da esquerda para a direita, as colunas representam: (1) representa a i-ésima parte, (2) $|L_i|$, (3) quantidade de estados s diferentes que compõem a i-ésima parte. As colunas 4, 5 e 6 fazem referência ao estados mais frequente na i-ésima parte: (4) estado s, $s \in \mathcal{S}$, dentre os $(\cdot, s) \in L_i$ que compõem a maior quantidade de elementos da parte L_i (5) frequência absoluta (#) do estado s, $s \in \mathcal{S}$, dentre os $(\cdot, s) \in L_i$ que compõem a maior quantidade de elementos da parte L_i e (6) frequência relativa (%) do estado s, $s \in \mathcal{S}$, dentre os $(\cdot, s) \in L_i$ que compõem a maior quantidade de elementos da parte L_i, em relação a $|L_i|$. (7) quantidade de sequências que fornecem elementos para a i-ésima parte. **Nota**: (i) o estado marcado com * encontra-se em mais de 50% e menos de 75% dos elementos da i-ésima partição, (ii) o estado marcado com ** faz parte de 75%, ou mais, dos elementos da i-ésima partição.

L_i	$\lvert L_i \rvert$	Quantidade de estados na parte L_i	Estado mais frequente na parte L_i			Quantidade de sequencias existente na parte L_i
			s	#	%	
1	58	3	aaa**	44	0,759	44
2	134	10	aac	41	0,306	44
3	93	7	tca	43	0,462	44
4	46	4	gtc**	43	0,935	43
5	177	11	cac	42	0,237	44
6	88	4	aag	43	0,489	44
7	41	4	cgg**	38	0,927	40
8	138	6	aat	44	0,319	44
9	45	3	aca**	40	0,889	41
10	46	4	tga**	43	0,935	44
11	127	6	cca	42	0,331	44
12	119	7	tta	41	0,345	43
13	84	8	ata	41	0,488	43
14	50	5	caa**	39	0,780	42
15	91	6	tgg	43	0,473	44
16	75	5	gca*	42	0,560	42
17	93	6	act	43	0,462	44
18	111	6	att	39	0,351	44
19	90	4	aga	44	0,489	44

Um detalhamento maior das partes do conjunto de partições obtidos para as sequências brasileiras é apresentado nas Tabelas B.5, B.6, B.7 e B.8.

Nas Tabelas B.5 e B.6 apresentamos quais são as partes L_k que um determinado

Tabela 35 – 2 de 2 - Análise descritiva das partes do modelo de partições estimado para as 44 sequências brasileiras de vírus da Zika. Da esquerda para a direita, as colunas representam: (1) representa a i-ésima parte, (2) $|L_i|$, (3) quantidade de estados s diferentes que compõem a i-ésima parte. As colunas 4, 5 e 6 fazem referência ao estados mais frequente na i-ésima parte: (4) estado s, $s \in \mathcal{S}$, dentre os $(\cdot, s) \in L_i$ que compõem a maior quantidade de elementos da parte L_i (5) frequência absoluta (#) do estado s, $s \in \mathcal{S}$, dentre os $(\cdot, s) \in L_i$ que compõem a maior quantidade de elementos da parte L_i e (6) frequência relativa (%) do estado s, $s \in \mathcal{S}$, dentre os $(\cdot, s) \in L_i$ que compõem a maior quantidade de elementos da parte L_i, em relação a $|L_i|$. (7) quantidade de sequências que fornecem elementos para a i-ésima parte. **Nota**: (i) o estado marcado com * encontra-se em mais de 50% e menos de 75% dos elementos da i-ésima partição, (ii) o estado marcado com ** faz parte de 75%, ou mais, dos elementos da i-ésima partição.

L_i	$\lvert L_i \rvert$	Quantidade de estados na parte L_i	Estado mais frequente na parte L_i			Quantidade de sequencias existente na parte L_i
			s	#	%	
20	85	4	agc*	43	0,506	44
21	47	4	gag**	44	0,936	44
22	98	5	cct	44	0,449	44
23	50	3	ttt**	44	0,880	44
24	45	2	cag**	44	0,978	44
25	42	1	atc**	42	1,000	42
26	53	5	cat**	43	0,811	44
27	84	4	ctc*	42	0,500	42
28	39	2	ccg**	38	0,974	39
29	128	5	gtt	44	0,344	44
30	88	4	gat*	44	0,500	44
31	46	3	gct**	41	0,891	42
32	43	2	tat**	42	0,977	42
33	50	4	ggc**	44	0,880	44
34	44	3	gaa**	42	0,955	42
35	43	1	gga**	43	1,000	43
36	42	2	gta**	41	0,976	41
37	41	2	tgc**	40	0,976	40
38	42	1	gtg**	42	1,000	42

estado s está presente. Além disso, temos a informação da quantidade de sequências que forneceram um determinado s para uma mesma parte L_k. Por exemplo, na Tabela B.5 o estado *aca* está presente em 4 (partes) a saber: (i) L_9 (ii) L_{11}, (iii) L_{10} e (iv) L_{12}. Além disso, a parte L_9 possui 40 elementos que possuem o estado *aca* na sua composição, já a parte L_{11}, 2 elementos com o estado *aca*, enquanto que as partes L_{10} e L_{12} possuem apenas um dos seus elementos contendo o estado *aca*.

Já nas Tabelas B.7 e B.8 apontamos quais e quantos são os estados que compõem uma parte L_k. Por exemplo, na Tabela B.7, temos que a parte L_{18} é composta por 6 (seis) estados distindos, sendo eles: *att*, *cgt*, *agt*, *act*, *cat* e *cga*. Ainda em relação a parte L_{18}, temos que os estados *att*, *cgt*, *agt*, *act*, *cat* e *cga*, estão presentes em 39, 37, 32, 1, 1 e 1 elementos de L_{18}, respectivamente. Detalhes adicionais dos resultados obtidos, quando do ajuste do modelo de partições para as sequências brasileiras, encontram-se disponíveis em <http://www.ime.unicamp.br/~jg/mpmmzika/MultPMM_zika_Brazil.txt>.

3.4.2.1.2 Todas as sequências:

Assim como foi feito apenas para as sequências brasileiras, aplicamos também a métrica $d((i,s),(j,r))$ para as 153 sequências provenientes de várias países. Para tal, empregando o método descrito na Seção C, para todos os elementos (i,s), (j,r) em $M = \{1,...,p\} \times \mathcal{S} = \{1,...,153\} \times \{a,c,g,t\}^3$. O conjunto $\{1,...,153\}$ faz referência as 153 sequências de ZV sequências vindas de vários países (Brasil, EUA, etc), de acordo com as Tabelas B.9 e B.10. Ao final do número de acesso de cada uma das 153 sequências, constam 3 (letras) que identificam o país de origem da sequência. Por exemplo, sequência número 1 = KX197192.1.BRA é brasileira, sequência número 2 = KX702400.1.VEN é venezuelana, ..., sequência número 153 = MF801426.1.NIC vem da Nicarágua. O ajuste do modelo de partições revelou um conjunto $\hat{\mathcal{L}}$, com $|\hat{\mathcal{L}}| = 55$. Assim, tendo as partes L_k do conjunto de partição $\hat{\mathcal{L}}$ sido determinadas, calculamos $N(L_k)$ e $N(L_k,x)$ onde $k = 1, 2, ..., 55$ e $x =$ a, c, g, t (Tabelas 36 e 37). Saber quanto valem $N(L_k)$ e $N(L_k,x)$ nos permite calcular $\hat{P}(x|L_k) = \dfrac{N(L_k,x)}{N(L_k)}$, $x =$ a, c, g, t (vide Tabelas 8 e 9 de (CORDEIRO et al., 2020)).

Com base nas distribuições de probabilidades $\hat{P}(x|L_k)$, $x =$ a, c, g, t, apresentadas nas Tabelas 8 e 9 de (CORDEIRO et al., 2020), do mesmo modo que aconteceu apenas para as sequências brasileiras, temos que a maioria das partes tendem de escolher os elementos a ou g como sendo o próximo elemento da cadeia.

O conjunto M, quanto todas as 153 sequências estão sendo consideradas, possui um total de $|M| = 153 \times 4^3 = 9792$ elementos o que resulta em $3 \times 9792 = 29376$ parâmetros a serem estimados, que no nosso caso, são as probabilidades de transição de cada elemento de M para algum dos elementos do alfabeto A. Mas, o modelo de partições possui $|\hat{\mathcal{L}}| = 55$ o que resulta em $3 \times 55 = 165$ parâmetros a serem estimados, ou seja, por volta de 0,6%

Tabela 36 – 1 de 2 - Valores de $N(L_i)$ e $N(L_i, x)$ onde $x = a, c, g, t$, obtidos após o ajuste do MMP para as 153 sequencias das américas.

L_i	$N(L_i)$	$N(L_i, a)$	$N(L_i, c)$	$N(L_i, g)$	$N(L_i, t)$
1	33307	11605	5995	11941	3766
2	51557	17167	8710	17093	8587
3	12851	4033	2710	4235	1873
4	38291	10970	6795	15121	5405
5	52693	20459	12228	8087	11919
6	21942	7375	6089	3172	5306
7	18021	7805	3299	2862	4055
8	22410	9265	5609	3169	4367
9	59268	22734	13890	7602	15042
10	34250	13399	6656	7787	6408
11	33751	12722	7168	9324	4537
12	14254	5694	3262	3220	2078
13	37115	13941	7152	10548	5474
14	11931	2620	2511	5255	1545
15	18843	3563	5296	7169	2815
16	37440	8639	6749	16574	5478
17	42067	9131	8377	18018	6541
18	30220	8147	7176	9404	5493
19	31516	7544	5828	12575	5569
20	32754	8938	7592	8512	7712
21	70954	18596	10811	25939	15608
22	31445	7436	7933	7655	8421
23	25469	4967	6404	7758	6340
24	30452	9089	7569	6464	7330
25	29283	8374	5838	6386	8685
26	27090	7710	5385	9054	4941
27	11913	2882	3107	3482	2442
28	30443	8284	9083	7577	5499

do número de parâmetros a serem estimados sem a estrutura de partições. Assim, em termos proporcionais, o modelo de partições para todas as 153 sequências é ainda mais parcimonioso que aquele obtido apenas para as sequências brasileiras.

As Tabelas 38 e 39 fornecem uma descrição das partes obtidas de todas as 153 sequências, após o ajuste do modelo de partições. Por exemplo, na Tabela 38 temos que a parte L_{14} é composta por 98 elementos e, apenas 5 (cinco) estados (dentre os 64 disponíveis) fazem parte da composição dos elementos de L_{14}. Além disso, o estado *aat* está presente em 88 dos 98 elementos (89,8%) de L_{14}, ou seja, 89,8% dos elementos de L_{14} contém o estado *aat* em sua composição. Por fim, temos que 95 (dentre as 153 sequências) contribuem com elementos para a partição L_{14}. A Tabela 39 fornece informação análoga a exibida pela Tabela 38, sendo a única diferença entre as duas é que a Tabela 38 descreve as partições de L_1 até L_{28}, enquanto que Tabela 39, de L_{29} até L_{55}.

Tabela 37 – 2 de 2 - Valores de $N(L_i)$ e $N(L_i, x)$ onde $x = a, c, g, t$, obtidos após o ajuste do MMP para as 153 sequencias das américas.

L_i	$N(L_i)$	$N(L_i, a)$	$N(L_i, c)$	$N(L_i, g)$	$N(L_i, t)$
29	55813	20691	11325	13476	10321
30	13052	2644	3463	4088	2857
31	41355	14537	6652	11410	8756
32	11357	1735	3999	2845	2778
33	19191	2915	6023	7212	3041
34	22176	2607	4595	8657	6317
35	17043	3887	3954	3959	5243
36	34805	6604	7877	12007	8317
37	22069	2937	4033	9739	5360
38	41322	5616	9031	16932	9743
39	15193	1356	3083	7006	3748
40	22964	2750	5140	8296	6778
41	24530	7525	6620	3284	7101
42	27881	7840	7843	3203	8995
43	19584	6668	5519	1964	5433
44	14897	5339	3325	1988	4245
45	13072	4106	2876	2323	3767
46	20897	3342	3548	9062	4945
47	27143	3218	7726	11778	4421
48	9339	1497	2634	2698	2510
49	19202	8322	3831	1153	5896
50	12034	4031	2837	1316	3850
51	9556	1931	3259	1866	2500
52	23260	5541	6525	2673	8521
53	22295	2623	5349	10442	3881
54	11509	2055	1419	5906	2129
55	29276	4555	6206	13424	5091

Mais detalhes a respeito das partes do conjunto de partições obtidos para todas as 153 sequências são apresentados nas Tabelas B.11, B.12, B.13 e B.14.

As partes L_k em que um determinado estado s está presente são apontadas nas Tabelas B.11 e B.12. Adicionalmente, também informam quantas sequências fornecem um determinado s para uma mesma parte L_k. Por exemplo, na Tabela B.11 o estado atg está presente em apenas 3 (três) partes a saber: (i) L_{16} (ii) L_{17} e (iii) L_{14}. Também, a parte L_{16} possui 140 elementos que possuem o estado atg na sua composição, já a parte L_{17}, 12 elementos com o estado atg, enquanto que a parte L_{14} possui apenas 1 elemento contendo o estado atg.

As Tabelas B.13 e B.14 indicam quais são os estados (e sua frequência absoluta) que compõem uma parte L_k. Por exemplo, na Tabela B.13, temos que a parte L_3 é composta de 5 (cinco) estados distindos, sendo eles: tag, aaa, gca, gga e gaa. Ainda para a parte L_3, temos que os estados tag, aaa, gca, gga e gaa, estão presentes em

Tabela 38 – 1 de 2 - Análise descritiva das partes do modelo de partições estimado para as 153 sequências de vírus da Zika. Da esquerda para a direita, as colunas representam: (1) representa a i-ésima parte, (2) quantidade de elementos existentes na i-ésima parte, (3) quantidade de estados s diferentes que compõem a i-ésima parte. As colunas 4, 5 e 6 fazem referência ao estados mais frequente na i-ésima parte: (4) estado s, (5) frequência absoluta (#) e (6) frequência relativa (%). (7) quantidade de sequências que fornecem elementos para a i-ésima parte. **Nota**: (i) o estado marcado com * encontra-se em mais de 50% e menos de 75% dos elementos da i-ésima partição, (ii) o estado marcado com ** faz parte de 75%, ou mais, dos elementos da i-ésima partição.

L_i	$\lvert L_i\rvert$	Quantidade de estados na parte L_i	Estado mais frequente na parte L_i			Quantidade de sequencias existente na parte L_i
			s	#	%	
1	155	3	aaa**	148	0,955	151
2	159	7	gga**	147	0,925	149
3	136	5	tag**	129	0,949	134
4	152	5	gaa**	145	0,954	147
5	363	10	aac	137	0,377	146
6	171	15	ccc**	137	0,801	147
7	158	6	gtc**	145	0,918	149
8	173	10	acc**	136	0,786	142
9	393	12	gac	140	0,356	152
10	156	8	agg**	135	0,865	140
11	136	4	aag**	121	0,890	126
12	161	5	cgg**	145	0,901	150
13	165	6	ggg**	133	0,806	139
14	98	5	aat**	88	0,898	95
15	273	6	cga	132	0,484	149
16	179	7	atg**	140	0,782	145
17	211	13	ctg*	143	0,678	152
18	151	6	caa**	138	0,914	144
19	151	8	gtg**	141	0,934	142
20	161	9	aca**	138	0,857	144
21	310	11	ttg	152	0,490	153
22	182	10	cca**	141	0,775	148
23	225	6	tct*	121	0,538	134
24	149	6	tga**	141	0,946	143
25	170	8	tca**	146	0,859	150
26	163	10	gca**	143	0,877	149
27	143	9	acg**	119	0,832	129
28	158	7	cag**	146	0,924	149

Tabela 39 – 2 de 2 - Análise descritiva das partes do modelo de partições estimado para as 153 sequências de vírus da Zika. Da esquerda para a direita, as colunas representam: (1) representa a i-ésima parte, (2) quantidade de elementos existentes na i-ésima parte, (3) quantidade de estados s diferentes que compõem a i-ésima parte. As colunas 4, 5 e 6 fazem referência ao estados mais frequente na i-ésima parte: (4) estado s, (5) frequência absoluta (#) e (6) frequência relativa (%). (7) quantidade de sequências que fornecem elementos para a i-ésima parte. **Nota**: (i) o estado marcado com * encontra-se em mais de 50% e menos de 75% dos elementos da i-ésima partição, (ii) o estado marcado com ** faz parte de 75%, ou mais, dos elementos da i-ésima partição.

L_i	$\lvert L_i\rvert$	Quantidade de estados na parte L_i	Estado mais frequente na parte L_i			Quantidade de sequencias existente na parte L_i
			s	#	%	
29	163	5	tgg**	143	0,877	149
30	128	9	ata**	109	0,852	118
31	150	5	gag**	146	0,973	148
32	153	6	ccg**	138	0,902	142
33	178	9	gtt**	141	0,792	148
34	154	7	act**	134	0,870	138
35	280	9	tcg	130	0,464	151
36	222	12	cat*	142	0,640	149
37	173	9	ctt*	129	0,746	142
38	388	12	att	132	0,340	143
39	149	5	ttt**	142	0,953	145
40	154	8	ggt**	138	0,896	141
41	137	7	agc**	129	0,942	134
42	165	8	ggc**	146	0,885	150
43	162	6	ctc**	129	0,796	143
44	139	10	tcc**	120	0,863	128
45	153	8	cgc**	137	0,895	145
46	176	9	cct**	138	0,784	147
47	165	9	gat**	143	0,867	147
48	150	7	tta**	132	0,880	140
49	154	3	atc**	152	0,987	152
50	133	6	tac**	120	0,902	127
51	149	8	gta**	141	0,946	145
52	153	7	tgc**	146	0,954	148
53	150	7	tgt**	141	0,940	144
54	148	5	tat**	142	0,959	143
55	164	8	gct**	143	0,872	148

129, 2, 2, 2 e 1 elementos de L_3, respectivamente. Para mais detalhes dos resultados obtidos, quando do ajuste do modelo de partições para todas as 153 sequências, consultar <http://www.ime.unicamp.br/~jg/mpmmzika/MultPMM_zika_All.txt>.

Nesta aplicação, temos, a rigor, dois conjuntos M: um considerando as 44 sequências de ZV brasileiras e, outro para as 153 sequências dos demais países (incluindo as 44 sequências do primeiro conjunto). Ambos os M foram particionados por meio do ajuste de MMP, utilizando a métrica $d((i,s),(j,r))$. O tamanho de ambos os conjuntos é bastante diferente, no primeiro caso, $|M| = 2816$ e, no segundo, $|M| = 9792$. Assim, inevitavelmente, as partes do MMP obtidas para as 44 sequências brasileiras serão diferentes daquelas obtidas para as 153 sequências dos demais países (que inclui as 44 brasileiras). A Tabela 40 fornece um resumo desta redistribuição. Por exemplo, os elementos que compõem a parte L_{19} obtida do ajuste do MMP às 44 sequências brasileiras, foram todos realocados na parte L_{21} obtida do ajuste do MMP as 153 sequências. Já a parte L_8 obtida do ajuste do MMP às 44 sequências brasileiras, teve seus elementos realocados entre as partes L_{16}, L_{17}, L_{14} e L_{55} obtida do ajuste do MMP as 153 sequências.

3.4.2.2 Conclusão

Em (CORDEIRO et al., 2020) utilizamos para a aplicação do método proposto, ou seja, o uso da métrica $d((i,s),(j,r))$ para ajustar um MMP (Definição 7), dois conjuntos de sequências genômicas de ZV: (1) uma composta por 44 sequências brasileiras e (2) 153 sequências de ZV provenientes de diversos países, dentre as quais estão incluídas as 44 brasileiras. Modelamos inicialmente o perfil estocástico das 44 sequências brasileiras através do ajuste de um MMP, onde obtivemos uma partição composta de 38 partes. Já para as 153 sequências, fizemos o mesmo, e encontramos uma partição de M composta de 55 partes. Ambos os modelos descrevem a transição que ocorre entre triplas compostas pelas bases $\{a,c,g,t\}$ para outro elemento qualquer do alfabeto genômico $A = \{a,c,g,t\}$. Os modelos obtidos são muito mais parcimoniosos, em comparação a uma modelagem sem a estrutura de partição do conjunto M e, proporcionalmente falando, o ganho com redução de parâmetros a serem estimados foi maior no caso do ajuste feito as 153 sequências de todos os países. Por fim, os MMP obtidos poderiam ser utilizados para comparação com uma sequência de ZV não utilizada na modelagem, de modo a, via uso da métrica $d((i,s),(j,r))$, identificar uma provável variante do ZV.

3.5 Conclusão do Capítulo 3

Neste capítulo, uma noção de equivalência foi definida (Definição 5) e, que uma partição de M, definida por esta equivalência é dita ser uma partição Markoviana (Definição 6). Quanto mais probabilidades de transição "redundantes" houver entre os

Tabela 40 – Realocação dos elementos pertencentes as 38 partes do MMP ajustado às 44 sequências brasileiras de ZV, entre as 55 partes obtidas do ajuste do MMP para as 153 sequências dos demais países (incluindo as 44 sequências brasileiras). L_k^B: partes do MMP para as 44 sequências brasileiras, L^F: partes do MMP para as 153 sequências dos demais países.

L_k^B	L^F
L_1^B	L_1^F (43), L_3^F (11), L_4^F (4)
L_2^B	L_5^F (77), L_8^F (49), L_6^F (3), L_9^F (3), L_{12}^F (1), L_{49}^F (1)
L_3^B	L_{25}^F (45), L_{45}^F (40), L_{10}^F (3), L_{35}^F (3), L_9^F (2)
L_4^B	L_7^F (45), L_{10}^F (1)
L_5^B	L_9^F (118), L_{44}^F (40), L_5^F (17), L_{49}^F (2)
L_6^B	L_{13}^F (51), L_{11}^F (35), L_{12}^F (2)
L_7^B	L_{12}^F (41)
L_8^B	L_{16}^F (58), L_{17}^F (56), L_{14}^F (22), L_{55}^F (2)
L_9^B	L_{20}^F (42), L_{22}^F (2), L_{25}^F (1)
L_{10}^B	L_{24}^F (40), L_{45}^F (6)
L_{11}^B	L_{35}^F (75), L_{22}^F (52)
L_{12}^B	L_{23}^F (71), L_{48}^F (40), L_{32}^F (5), L_{35}^F (2), L_{22}^F (1)
L_{13}^B	L_{27}^F (42), L_{30}^F (25), L_{23}^F (12), L_{20}^F (2), L_{26}^F (2), L_{22}^F (1)
L_{14}^B	L_{18}^F (39), L_{20}^F (6), L_{27}^F (3), L_{28}^F (2)
L_{15}^B	L_{29}^F (53), L_{10}^F (36), L_{12}^F (1), L_{31}^F (1)
L_{16}^B	L_{26}^F (44), L_3^F (29), L_{18}^F (2)
L_{17}^B	L_{34}^F (43), L_{40}^F (43), L_{35}^F (2), L_{36}^F (2), L_{39}^F (2), L_{48}^F (1)
L_{18}^B	L_{38}^F (95), L_{37}^F (5), L_{55}^F (5), L_{17}^F (3), L_{36}^F (2), L_{19}^F (1)
L_{19}^B	L_{21}^F (90)
L_{20}^B	L_6^F (42), L_{41}^F (40), L_{43}^F (2), L_{42}^F (1)
L_{21}^B	L_{31}^F (43), L_{10}^F (4)
L_{22}^B	L_{46}^F (60), L_{37}^F (36), L_{38}^F (2)
L_{23}^B	L_{39}^F (43), L_{37}^F (7)
L_{24}^B	L_{28}^F (45)
L_{25}^B	L_{49}^F (42)
L_{26}^B	L_{36}^F (52), L_{27}^F (1)
L_{27}^B	L_{43}^F (46), L_{50}^F (36), L_9^F (2)
L_{28}^B	L_{32}^F (39)
L_{29}^B	L_{15}^F (75), L_{33}^F (52), L_{17}^F (1)
L_{30}^B	L_{47}^F (46), L_{53}^F (38), L_{38}^F (2), L_{40}^F (1), L_{55}^F (1)
L_{31}^B	L_{55}^F (44), L_{47}^F (1), L_{53}^F (1)
L_{32}^B	L_{54}^F (43)
L_{33}^B	L_{42}^F (49), L_{43}^F (1)
L_{34}^B	L_4^F (42), L_2^F (1), L_{19}^F (1)
L_{35}^B	L_2^F (43)
L_{36}^B	L_{51}^F (42)
L_{37}^B	L_{52}^F (41)
L_{38}^B	L_{19}^F (42)

elementos de $\mathcal{F}$, menor será a cardinalidade da partição obtida de M.

Na seção 3.2 a métrica $d((i,s),(j,r))$ é definida (tal como o processo de sua obtenção através do BIC). No Exemplo 7 apresentado, criado para ilustrar o cálculo de $d((i,s),(j,r))$, observa-se que os valores de $d((i,s),(j,r))$ entre elementos pertencentes a uma mesma partição são próximos de zero.

Já a seção 3.3 trouxe resultados teóricos da métrica $d((i,s),(j,r))$. Aqui, vale destaque os Teoremas 9, 10, 11 e 12. O Teorema 9 mostrou que o critério BIC indica que $(i,s) \sim (j,r)$ se, e somente se, $d((i,s),(j,r)) < 1$, i.e., a divergência $d((i,s),(j,r))$ está relacionada ao critério BIC. O Teorema 10 nos disse que $d((i,s),(j,r))$ é estatisticamente consistente, pois, para um n_{min} suficientemente grande, ela é capas de detectar se as leis sob comparação são discrepantes ou semelhantes. O Teorema 11 provou que $d((i,s),(j,r))$ é, de fato, uma métrica definida em M, ou seja, é não-negativa, simétrica e obedece a desigualdade triangular. Assim, M é um espaço métrico. E, por fim, o Teorema 12 mostrou que a partição verdadeira $\mathcal{L}$, obtida da Definição 6, de M, pode ser estimada eventualmente quase certamente, por meio do BIC, ou seja, para um n_{min} suficientemente grande, a partição verdadeira $\mathcal{L}$ pode ser recuperada pelo BIC.

A seção 3.4 trouxe dois estudos de caso, já publicado, contendo uma aplicação utilizando dados reais da métrica $d((i,s),(j,r))$ juntamente com modelo de Markov de partições (MMP). Na subseção 3.4.1, o trabalho (CORDEIRO et al., 2019b) (vide apêndice) foi abordado, onde o conhecimento prévio existente entre as sequências foi utilizado de modo a, explicar-se o comportamento probabilístico das sequências por um modelo mais parcimonioso. Finalmente, na subseção 3.4.2, a aplicação feita em (CORDEIRO et al., 2020) foi apresentada. Em (CORDEIRO et al., 2020), a métrica $d((i,s),(j,r))$ juntamente com modelo de Markov de partições (MMP) foram usadas para obtenção de um perfil para o ZV.

4 Conclusão

Este trabalho foi desenvolvido tendo em vista dois objetivos principais: (i) a utilização da métrica d_s, via $\delta(i)$, como um classificador (ranqueador) de uma coleção de realizações de processos estocásticos e, (ii) desenvolvimento da métrica $d((i,s),(j,r))$, extensão de d_s, e, como esta pode ser empregada na estimação do modelo de Markov de Partições.

No Capítulo 2 o processo de construção da métrica d_s foi apresentado e, como esta pode ser usada na classificação de uma coleção de amostras de processos estocásticos, via $\delta(i)$. Nas Subseções 2.4.1 e 2.4.2 mostramos duas aplicações a problemas reais. Nestas aplicações, dados genômicos dos vírus da dengue e da Zika são tratados como realizações de processos estocásticos markovianos, assumindo valores num alfabeto $A = \{a, g, t, c\}$ e tendo uma ordem (memória) o finita e, assim, classificados. Essa classificação se deu pela representatividade da sequência em relação a lei predominante do conjunto destas sequências. Em ambos os casos, indicamos qual(is) seria(m) as amostras mais e menos representativas destes conjuntos. Como já dito, as sequências mais representativas destes conjuntos poderiam ser utilizadas como uma espécie de "padrão-ouro", ou seja, tais sequências podem ser utilizadas como um padrão a ser comparada com sequências futuras ou provenientes de outros coleções.

No entanto, a abordagem de obtenção de uma sequência padrão, ou seja, a mais representativa de um conjunto, pelo uso de d_s, via $\delta(i)$, quando foi desenvolvida, tinha o propósito apenas de classificar e não de definir um modelo conjunto. Assim sendo, a informação contida nas demais sequências acaba sendo pura e simplesmente desprezada, uma vez que apenas uma sequência é considerada. Uma forma de contornar esta característica foi apresentada no Capítulo 3.

No Capítulo 3, definimos uma noção de equivalência (Definição 5) entre elementos de um conjunto $M = \{1, 2, ..., p\} \times A^o$, onde a equivalência é dada pela redundância nas probabilidades de transição condicionada entre um estado s na i-ésima sequência (i, s) e, um estado r na j-ésima sequência (j, r) para qualquer um dos elementos do alfabeto A, no qual os processos estão definidos. Essa relação de equivalência nos permitiu particionar o conjunto M, onde, cada uma das partes desta partição, é composta apenas de elementos com as mesmas probabilidades de transição. Desta partição, podemos obter um modelo que descreve a lei de formação da coleção de realizações de processos estocásticos. Vimos que este modelo poderia ser estimado diretamente via BIC deste modelo, mas, isso seria inviável, uma vez que o conjunto de todas as partições possíveis de M é, em geral, muito grande. Assim, mostramos como a partição ótima, i.e., aquela que maximiza o BIC, pode

ser obtida através da métrica $d((i,s),(j,r))$ (Seção C).

Ainda no Capítulo 3, ainda mostramos algumas das propriedades teóricas conhecidas da métrica $d((i,s),(j,r))$. Dentre elas, destacamos o resultado que garante que $d((i,s),(j,r))$ é, de fato uma métrica, definida no espaço M, ou seja, M é um espaço métrico. Também vale enfatizar o resultado que afirma $d((i,s),(j,r))$ ser estatisticamente consistência na determinação da equivalência de (i,s) e (j,r). Nos dois estudos de caso apresentados, publicados em (CORDEIRO et al., 2019b; CORDEIRO et al., 2020) obtivemos um modelo de Markov de partições, usando $d((i,s),(j,r))$ de coleções de dados genômicos do vírus EBV (CORDEIRO et al., 2019b) e da Zika (CORDEIRO et al., 2020). Em ambas as análises, obtivemos um modelo que descreve o perfil estocástico de ambos os conjuntos de sequências. Na duas situações, o modelo obtido é mais parcimonioso que aquele a ser obtido sem que a estrutura de partições tivesse sido considerada. Tanto em (CORDEIRO et al., 2019b) quanto em (CORDEIRO et al., 2020) poder-se-ía utilizar o modelo obtido para comparações futuras com outras sequências provenientes de outras localidades ou de períodos de tempos diferentes. Esta comparação poderia, por exemplo, apontar possíveis diferenças existentes numa determinada cepa, em comparação a um padrão pré-estabelecido.

Como trabalhos futuros, sugerimos o desenvolvimento de um teste de hipóteses onde a hipótese nula refere-se a igualdade de $P^i(\cdot|s) = P^j(\cdot|s)$ sendo a métrica $d((i,s),(j,r))$ a estatística do teste.

Referências

CORDEIRO, M. T. A.; GARCÍA, J. E.; GONZÁLEZ-LÓPEZ, V. A.; LONDOÑO, S. L. M. Classification of autochthonous dengue virus type 1 strains circulating in japan in 2014. *4open*, v. 2, p. 20, 2019. Disponível em: <https://doi.org/10.1051/fopen/2019018>. Citado 13 vezes nas páginas 7, 8, 9, 12, 25, 26, 28, 58, 60, 62, 63, 64 e 65.

_____. Stochastic profile of epstein-barr virus in nasopharyngeal carcinoma settings. *4open*, v. 2, p. 25, 2019. Disponível em: <https://doi.org/10.1051/fopen/2019020>. Citado 9 vezes nas páginas 7, 8, 27, 68, 81, 82, 83, 96 e 98.

_____. Partition markov model for multiple processes. *Mathematical Methods in the Applied Sciences*, n/a, n. n/a, 2020. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/mma.6079>. Citado 15 vezes nas páginas 7, 8, 26, 27, 68, 78, 79, 80, 81, 84, 85, 89, 94, 96 e 98.

CSISZÁR, I.; SHIELDS, P. C. The consistency of the BIC Markov order estimator. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 28, n. 6, p. 1601 – 1619, 2000. Disponível em: <https://doi.org/10.1214/aos/1015957472>. Citado na página 30.

CSISZÁR, I.; TALATA, Z. Context tree estimation for not necessarily finite memory processes, via bic and mdl. *IEEE Transactions on Information Theory*, v. 52, n. 3, p. 1007–1016, March 2006. ISSN 1557-9654. Citado 2 vezes nas páginas 30 e 79.

FERNÁNDEZ, M.; GARCÍA, J. E.; GHOLIZADEH, R.; GONZÁLEZ-LÓPEZ, V. A. Sample selection procedure in daily trading volume processes. *Mathematical Methods in the Applied Sciences*, v. 0, n. 0, 2019. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/mma.5705>. Citado 9 vezes nas páginas 7, 8, 25, 41, 42, 44, 58, 59 e 60.

GARCÍA, J. E.; GHOLIZADEH, R.; GONZÁLEZ-LÓPEZ, V. A. Linguistic compositions highly volatile in portuguese. *Cadernos de Estudos Linguísticos*, v. 59, n. 3, p. 617–630, 12 2017. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/cel/article/view/8651002>. Citado na página 25.

_____. A bic-based consistent metric between markovian processes. *Applied Stochastic Models in Business and Industry*, v. 34, n. 6, p. 868–878, 2018. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.2346>. Citado 9 vezes nas páginas 7, 8, 24, 28, 33, 35, 36, 58 e 60.

_____. Stochastic distance between burkitt lymphoma/leukemia strains. In: _____. *Demography and Health Issues: Population Aging, Mortality and Data Analysis*. Cham: Springer International Publishing, 2018. p. 143–153. ISBN 978-3-319-76002-5. Disponível em: <https://doi.org/10.1007/978-3-319-76002-5_13>. Citado 2 vezes nas páginas 25 e 60.

GARCÍA, J. E.; GONZÁLEZ-LÓPEZ, V. A. Consistent estimation of partition markov models. *Entropy*, v. 19, n. 4, 2017. ISSN 1099-4300. Disponível em: <https://www.mdpi.com/1099-4300/19/4/160>. Citado 2 vezes nas páginas 30 e 81.

GARCÍA, J. E.; GONZÁLEZ-LÓPEZ, V. A.; ANDRADE, F. K. de. Dissimilarity between markovian processes applied to industrial processes. In: AIP PUBLISHING LLC. *AIP Conference Proceedings*. [S.l.], 2017. v. 1863, n. 1, p. 220002. Citado na página 25.

GARCÍA, J. E.; GONZÁLEZ-LÓPEZ, V. A.; LONDOÑO, S. L. M.; CORDEIRO, M. T. A. Similarity between strains of zika from tropical and subtropical regions. *AA*, 2018. Citado 24 vezes nas páginas 7, 8, 10, 11, 12, 25, 26, 27, 28, 43, 44, 45, 47, 48, 49, 51, 52, 53, 54, 55, 56, 57, 84 e 85.

KWOK, H.; TONG, A. H.; LIN, C. H.; LOK, S.; FARRELL, P. J.; KWONG, D. L.; CHIANG, A. K. Genomic sequencing and comparative analysis of epstein-barr virus genome isolated from primary nasopharyngeal carcinoma biopsy. *PloS one*, Public Library of Science, v. 7, n. 5, p. e36939, 2012. Citado 2 vezes nas páginas 60 e 83.

MUSTAFA, M.; RASOTGI, V.; JAIN, S.; GUPTA, V. Discovery of fifth serotype of dengue virus (denv-5): A new public health dilemma in dengue control. *Medical Journal Armed Forces India*, v. 71, n. 1, p. 67–70, 2015. ISSN 0377-1237. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0377123714001725>. Citado na página 59.

SCHWARZ, G. Estimating the Dimension of a Model. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461 – 464, 1978. Disponível em: <https://doi.org/10.1214/aos/1176344136>. Citado 2 vezes nas páginas 31 e 44.

TAJIMA, S.; NAKAYAMA, E.; KOTAKI, A.; MOI, M. L.; IKEDA, M.; YAGASAKI, K.; SAITO, Y.; SHIBASAKI, K. ichi; SAIJO, M.; TAKASAKI, T. Whole genome sequencing–based molecular epidemiologic analysis of autochthonous dengue virus type 1 strains circulating in japan in 2014. *Japanese journal of infectious diseases*, National Institute of Infectious Diseases, Japanese Journal of Infectious ..., p. JJID–2016, 2016. Citado 4 vezes nas páginas 59, 60, 61 e 66.

ZENG, M.-S.; DA-JIANGLI; LIU, Q.-L.; LI-BINGSONG; LI, M.-Z.; ZHANG, R.-H.; YU, X.-J.; WANG, H.-M.; ERNBERG, I.; ZENG, Y.-X. Genomic sequence analysis of epstein-barr virus strain gd1 from a nasopharyngeal carcinoma patient. *Journal of virology*, Am Soc Microbiol, v. 79, n. 24, p. 15323–15330, 2005. Citado na página 60.

APÊNDICE A – Tabelas e gráficos do Capítulo 2

Tabela A.1 – Sequência $x_{1,1}^{n_1}$ gerada de uma CM com ordem $o = 3$, $n_1 = 400$ e probabilidades de transição dada pela Tabela 5.

$x_{1,1}^{n_1} =$ 0000001110011110000000000101111111111000111100101111111111111011111000001111111111111111
1111100100000101111111111111111111101010010011000100101100111101010000000000001111111111111111
0010101010000000101110100101111111111111111110000000101110001001110111111111111110001111111101
1111000000000111111111000011010001001010011111111100111110101011001111100100111111111101000
1111111111100100111100101111100100100 1001

Tabela A.2 – Sequência $x_{2,1}^{n_2}$ gerada de uma CM com ordem $o = 3$, $n_2 = 401$ e probabilidades de transição dada pela Tabela 5.

$x_{2,1}^{n_2} =$ 1010110000101001110011111001011111111111110000001011111100001001001110111111110001011111
1111111111111101111111100000010110100010001111111110100110010000001111100011111100010011111
1011111111111111110011001111111111111101110110010000001011001011101111111110001011111001011
1110001000000000001011111011111111001001111111111111111111110010111111111111111111111000110
1111111111111100100100100000000010000001001

Tabela A.3 – Sequência $x_{3,1}^{n_3}$ gerada de uma CM com ordem $o = 3$, $n_3 = 402$ e probabilidades de transição dada pela Tabela 5.

$x_{3,1}^{n_3} =$ 0111001111111110111111111111111001111110010011111110100101110100101111111100111111111101
1110010000111111111111111111111111011100001110000001010101011100100111111111111100111111111111
1100101001111111111011011111100101111110010000000000000100000000010010100111100101110111111111
1111111111110010000000011100000010000000000001111111111111111011110010111111110001001111111 1
1101111111000001111111111111001001001000000 0

Tabela A.4 – Sequência $x_{4,1}^{n_4}$ gerada de uma CM com ordem $o = 3$, $n_4 = 403$ e probabilidades de transição dada pela Tabela 5

$x_{4,1}^{n_4} =$ 0010010111100111000000111111111101101001001111111111111111101111111111111111111100111
1001100000011111110001111111111111001111111111100001111110110010010010010111111111100111111
1111011111111111000000001001111100011111110000010011110111001111110011110000000111111100111
1111100000001010111111101101110101000000111111111000111110001111010111110011100111111111111
1101110100011001111111110011111111110111111010

Tabela A.5 – Sequência $x_{5,1}^{n_5}$ gerada de uma CM com ordem $o = 3$, $n_5 = 404$ e probabilidades de transição dada pela Tabela 6.

$x_{5,1}^{n_5} =$ 0101100101010010110110100001010101010110010110010110010010001100100111110000110011001
000101011001010110111100101011001001110010101010100100101100101010100110110001110001001000010
101001100110101100011001010100010001010101010010110000100001100110001010001011001100100101
010011001010110011110010110001100100001011000010101011111111110001010110000101101010010001001
010101000101011000100100010001010100010001000

Tabela A.6 – Sequência $x_{6,1}^{n_6}$ gerada de uma CM com ordem $o = 3$, $n_6 = 405$ e probabilidades de transição dada pela Tabela 6.

$x_{6,1}^{n_6} =$ 0000101010001000010001010101011100010010110000010101011000101100000101011001011000110
011001010101001000010110011001010100010001110010101100011001100101011001100110001010010
101010101011000110010110010101010101010111100001100010110010101010111100011100100010100101100
000110001011000001100001010101001010001111100100010001010100001011000011001010110001000100
100111100010001010100101010001010000100010010100

APÊNDICE B – Tabelas e gráficos do Capítulo 3

Tabela B.1 – Sequência $x_{1,1}^{1001}$ do Exemplo 7.

$x_{1,1}^{1001} =$ 0000001101101101010111101111111010111110111011100111111011101011110111101011101111
0000011101111011011111010111011111111101010101011111100111111011101001110111110100101111101
0100001111011110110111100011100011101100011101100110111110111111101011111010101111101111010
1111001110110011101111111101011110111000011101111000010111110111100001111111110111101101111
1011111000111111110111110110111100011111111111110111101010110101011110111111101001110110001 1
1111111111101011111010111101100001111010111011101011011111111101010101011111101011110101 11
0101111111111011011011001111111110111110110110111111111110111110111011100011110101110111011
1101111101010111000010110111101110111111111111111111110101010101110101111101011011011110001 0
1101101011101011111101100111011101011100011010011111111101111011110111011101111011010110110 1
0110111010111100101111101100000001110100000011100011011011010110101110111011001111111111110
1111100111100101111110111111111111101010111011111011111100111111010111010101111111010110110
1101111

Tabela B.2 – Sequência $x_{2,1}^{1002}$ do Exemplo 7.

$x_{2,1}^{1002} =$ 01111111101001111111111111101000111101111110010011101010100011011100100101111111111
01010011110011101111101001010000010101000111011111111110100010111100000010011011111001111000
00111111111000101000111010011111110011111111111111110101000000000100011111110000111001111001111
11111111110000001110110100001101101111110101100110011010001111101110001001011110011111111100
00011110101111110111110011111001011111111111111011000010011001101110011111111111111110100110
10110011101111111010001111111110100100100101110001011000101100101111111001010110011101001010
00100111010011010010110110101111111111111111110000101100011111101110110010001100011111101100
10100110100100100101111010111111111011111111111111111111111111111111111001100010111100001100100001
00111001111001000101111001111100110011111101111100101100111100101010010101111111100111100100
01001111101101101101111011001101111111101100101110101011110011010001001111001101001001010
00111111010011110010010100011101010001111111110001001111100101111111100110101111111010111111
01111001

Tabela B.5 – 1 de 2 - Análise descritiva das partes do modelo de partições estimado para as 44 sequências brasileiras de vírus da Zika. Da esquerda para a direita, as colunas representam: (1) s: estado s, (2) L_i (# seq): a quantidade entre parêntesis é o número de sequências que fornecem o estado s para a parte L_i. **Nota**: A soma dos números entre parêntesis, por linha, é igual a 44.

s	L_i (# seq)
aaa	L_1 (44).
aac	L_2 (41), L_3 (1), L_4 (1), L_5 (1).
aag	L_6 (43), L_7 (1).
aat	L_8 (44).
aca	L_9 (40), L_{11} (2), L_{10} (1), L_{12} (1).
acc	L_2 (41), L_5 (2), L_4 (1).
acg	L_{13} (29), L_{14} (7), L_9 (3), L_{10} (1), L_{15} (1), L_{16} (1), L_6 (1), L_7 (1).
act	L_{17} (43), L_{18} (1).
aga	L_{19} (44).
agc	L_{20} (43), L_3 (1).
agg	L_{15} (42), L_{21} (1), L_4 (1).
agt	L_{18} (32), L_{22} (11), L_{23} (1).
ata	L_{13} (41), L_{12} (1), L_{14} (1), L_{24} (1).
atc	L_{25} (42), L_2 (1), L_5 (1).
atg	L_8 (44).
att	L_{18} (39), L_{17} (3), L_{26} (2).
caa	L_{14} (39), L_{13} (4), L_{16} (1).
cac	L_5 (42), L_{27} (2).
cag	L_{24} (44).
cat	L_{26} (43), L_{18} (1).
cca	L_{11} (42), L_9 (2).
ccc	L_{20} (37), L_2 (6), L_5 (1).
ccg	L_{28} (38), L_{12} (4), L_{29} (1), L_{30} (1).
cct	L_{22} (44).
cga	L_{29} (39), L_8 (2), L_{13} (1), L_{18} (1), L_{30} (1).
cgc	L_3 (42), L_{21} (1), L_5 (1).
cgg	L_7 (38), L_6 (3), L_{15} (2), L_2 (1).
cgt	L_{18} (37), L_{31} (3), L_{22} (2), L_{14} (1), L_{32} (1).
cta	L_{12} (35), L_{13} (4), L_{11} (2), L_{26} (2), L_{17} (1).
ctc	L_{27} (42), L_{33} (1), L_5 (1).
ctg	L_8 (44).
ctt	L_{22} (39), L_{23} (5).

Tabela B.6 – 2 de 2 - Análise descritiva das partes do modelo de partições estimado para as 44 sequências brasileiras de vírus da Zika. Da esquerda para a direita, as colunas representam: (1) s: estado s, (2) L_i (# seq): a quantidade entre parêntesis é o número de sequências que fornecem o estado s para a parte L_i. **Nota**: A soma dos números entre parêntesis, por linha, é igual a 44.

s	L_i (# seq)
gaa	L_{34} (42), L_1 (1), L_{16} (1).
gac	L_5 (40), L_2 (4).
gag	L_{21} (44).
gat	L_{30} (44).
gca	L_{16} (42), L_{14} (2).
gcc	L_2 (36), L_5 (7), L_{20} (1).
gcg	L_{29} (43), L_{13} (1).
gct	L_{31} (41), L_8 (3).
gga	L_{35} (43), L_{19} (1).
ggc	L_{33} (44).
ggg	L_6 (41), L_{15} (2), L_7 (1).
ggt	L_{17} (43), L_{26} (1).
gta	L_{36} (41), L_{11} (1), L_{28} (1), L_{37} (1).
gtc	L_4 (43), L_2 (1).
gtg	L_{38} (42), L_{34} (1), L_8 (1).
gtt	L_{29} (44).
taa	L_{11} (39), L_3 (2), L_{12} (1), L_{13} (1), L_{19} (1).
tac	L_{27} (39), L_{20} (4), L_{33} (1).
tag	L_{16} (30), L_1 (13), L_{34} (1).
tat	L_{32} (42), L_{22} (2).
tca	L_3 (43), L_{10} (1).
tcc	L_5 (42), L_2 (1), L_{27} (1).
tcg	L_{11} (41), L_{17} (2), L_3 (1).
tct	L_{12} (36), L_{26} (5), L_{13} (3).
tga	L_{10} (43), L_{15} (1).
tgc	L_{37} (40), L_{33} (4).
tgg	L_{15} (43), L_{21} (1).
tgt	L_{30} (42), L_{31} (2).
tta	L_{12} (41), L_{17} (1), L_{29} (1), L_{36} (1).
ttc	L_5 (39), L_3 (3), L_2 (2).
ttg	L_{19} (44).
ttt	L_{23} (44).

Tabela B.7 – 1 de 2 - Análise descritiva das partes do modelo de partições estimado para as 44 sequências brasileiras de vírus da Zika. Da esquerda para a direita, as colunas representam: (1) L_i: i-ésima parte do modelo de partições, (2) s (freq): a quantidade entre parêntesis é o número de elementos de L_i que possuem o estado s em sua composição. **Nota**: A soma dos números entre parêntesis, por linha, é igual a $|L_i|$.

L_i	s (freq)
1	aaa (44), tag (13), gaa (1).
2	aac (41), acc (41), gcc (36), ccc (6), gac (4), ttc (2), atc (1), cgg (1), gtc (1), tcc (1).
3	tca (43), cgc (42), ttc (3), taa (2), aac (1), agc (1), tcg (1).
4	gtc (43), aac (1), acc (1), agg (1).
5	cac (42), tcc (42), gac (40), ttc (39), gcc (7), acc (2), aac (1), atc (1), ccc (1), cgc (1), ctc (1).
6	aag (43), ggg (41), cgg (3), acg (1).
7	cgg (38), aag (1), acg (1), ggg (1).
8	aat (44), atg (44), ctg (44), gct (3), cga (2), gtg (1).
9	aca (40), acg (3), cca (2).
10	tga (43), aca (1), acg (1), tca (1).
11	cca (42), tcg (41), taa (39), aca (2), cta (2), gta (1).
12	tta (41), tct (36), cta (35), ccg (4), aca (1), ata (1), taa (1).
13	ata (41), acg (29), caa (4), cta (4), tct (3), cga (1), gcg (1), taa (1).
14	caa (39), acg (7), gca (2), ata (1), cgt (1).
15	tgg (43), agg (42), cgg (2), ggg (2), acg (1), tga (1).
16	gca (42), tag (30), acg (1), caa (1), gaa (1).
17	act (43), ggt (43), att (3), tcg (2), cta (1), tta (1).
18	att (39), cgt (37), agt (32), act (1), cat (1), cga (1).
19	aga (44), ttg (44), gga (1), taa (1).

Tabela B.8 – 2 de 2 - Análise descritiva das partes do modelo de partições estimado para as 44 sequências brasileiras de vírus da Zika. Da esquerda para a direita, as colunas representam: (1) L_i: i-ésima parte do modelo de partições, (2) s (freq): a quantidade entre parêntesis é o número de elementos de L_i que possuem o estado s em sua composição. **Nota**: A soma dos números entre parêntesis, por linha, é igual a $|L_i|$.

L_i	s (freq)
20	agc (43), ccc (37), tac (4), gcc (1).
21	gag (44), agg (1), cgc (1), tgg (1).
22	cct (44), ctt (39), agt (11), cgt (2), tat (2).
23	ttt (44), ctt (5), agt (1).
24	cag (44), ata (1).
25	atc (42).
26	cat (43), tct (5), att (2), cta (2), ggt (1).
27	ctc (42), tac (39), cac (2), tcc (1).
28	ccg (38), gta (1).
29	gtt (44), gcg (43), cga (39), ccg (1), tta (1).
30	gat (44), tgt (42), ccg (1), cga (1).
31	gct (41), cgt (3), tgt (2).
32	tat (42), cgt (1).
33	ggc (44), tgc (4), ctc (1), tac (1).
34	gaa (42), gtg (1), tag (1).
35	gga (43).
36	gta (41), tta (1).
37	tgc (40), gta (1).
38	gtg (42).

Tabela B.9 – 1 de 2 - Sequências completas de ZV provenientes de países das Américas. As 3 (letras) últimas letras do número de acesso identificam o país de origem da sequência: (1) BRA: Brasil, (2) USA: Estados Unidos, (3) DOM: República Dominicana, (4) MEX: México, (5) HND: Honduras, (6) NIC: Nicarágua, (7) JAM: Jamaica, (8) COL: Colômbia, (9) PRI: Porto Rico , (10) VEN: Venezuela, (11) CUB: Cuba, (12) MTQ: Martinica.

i	número de acesso	i	número de acesso	i	número de acesso
1	KX197192.1.BRA	32	KY014319.2.HND	63	KY559010.1.BRA
2	KX702400.1.VEN	33	KY014320.2.BRA	64	KY559011.1.BRA
3	KX832731.1.USA	34	KY014321.2.DOM	65	KY559012.1.BRA
4	KX842449.2.USA	35	KY014325.2.USA	66	KY559013.1.BRA
5	KX893855.1.VEN	36	KY014326.1.USA	67	KY559014.1.BRA
6	KX922703.1.USA	37	KY014327.2.HND	68	KY559015.1.BRA
7	KX922704.1.USA	38	KY075932.1.USA	69	KY559017.1.BRA
8	KX922705.1.USA	39	KY075933.1.USA	70	KY559018.1.BRA
9	KX922706.1.USA	40	KY075934.1.USA	71	KY559019.1.BRA
10	KX922707.1.USA	41	KY075935.1.USA	72	KY559021.1.BRA
11	KY014295.2.USA	42	KY075936.1.USA	73	KY559023.1.BRA
12	KY014296.2.BRA	43	KY325464.1.USA	74	KY559024.1.BRA
13	KY014297.2.BRA	44	KY325465.1.USA	75	KY559027.1.BRA
14	KY014298.1.USA	45	KY325466.1.USA	76	KY559031.1.BRA
15	KY014300.2.DOM	46	KY325467.1.USA	77	KY559032.1.BRA
16	KY014301.2.BRA	47	KY325468.1.USA	78	KY785410.1.BRA
17	KY014302.3.DOM	48	KY325469.1.USA	79	KY785412.1.USA
18	KY014303.2.DOM	49	KY325471.1.USA	80	KY785413.1.DOM
19	KY014304.2.DOM	50	KY325472.1.USA	81	KY785414.1.HND
20	KY014305.2.DOM	51	KY325473.1.USA	82	KY785415.1.DOM
21	KY014306.2.HND	52	KY325476.1.USA	83	KY785417.1.COL
22	KY014307.2.BRA	53	KY325477.1.USA	84	KY785418.1.HND
23	KY014308.2.BRA	54	KY325479.1.USA	85	KY785419.1.JAM
24	KY014310.2.HND	55	KY558999.1.BRA	86	KY785420.1.DOM
25	KY014312.2.HND	56	KY559001.1.BRA	87	KY785423.1.DOM
26	KY014313.2.BRA	57	KY559003.1.BRA	88	KY785424.1.JAM
27	KY014314.2.DOM	58	KY559004.1.BRA	89	KY785426.1.BRA
28	KY014315.2.HND	59	KY559005.1.BRA	90	KY785427.1.BRA
29	KY014316.2.USA	60	KY559006.1.BRA	91	KY785429.1.BRA
30	KY014317.2.BRA	61	KY559007.1.BRA	92	KY785430.1.JAM
31	KY014318.3.DOM	62	KY559009.1.BRA	93	KY785432.1.JAM

Tabela B.10 – 2 de 2 - Sequências completas de ZV provenientes de países das Américas. As 3 (letras) últimas letras do número de acesso identificam o país de origem da sequência: (1) BRA: Brasil, (2) USA: Estados Unidos, (3) DOM: República Dominicana, (4) MEX: México, (5) HND: Honduras, (6) NIC: Nicarágua, (7) JAM: Jamaica, (8) COL: Colômbia, (9) PRI: Porto Rico , (10) VEN: Venezuela, (11) CUB: Cuba, (12) MTQ: Martinica.

i	número de acesso	i	número de acesso	i	número de acesso
64	KY559011.1.BRA	95	KY785435.1.DOM	126	KY817930.1.BRA
65	KY559012.1.BRA	96	KY785437.1.BRA	127	MF434516.1.NIC
66	KY559013.1.BRA	97	KY785439.1.BRA	128	MF434517.1.NIC
67	KY559014.1.BRA	98	KY785441.1.DOM	129	MF434518.1.NIC
68	KY559015.1.BRA	99	KY785442.1.HND	130	MF434520.1.NIC
69	KY559017.1.BRA	100	KY785444.1.HND	131	MF434521.1.NIC
70	KY559018.1.BRA	101	KY785445.1.USA	132	MF434522.1.NIC
71	KY559019.1.BRA	102	KY785447.1.DOM	133	MF438286.1.CUB
72	KY559021.1.BRA	103	KY785448.1.HND	134	MF801391.1.MEX
73	KY559023.1.BRA	104	KY785449.1.DOM	135	MF801395.1.MEX
74	KY559024.1.BRA	105	KY785450.1.BRA	136	MF801396.1.MEX
75	KY559027.1.BRA	106	KY785451.1.MTQ	137	MF801398.1.MEX
76	KY559031.1.BRA	107	KY785452.1.HND	138	MF801402.1.MEX
77	KY559032.1.BRA	108	KY785453.1.DOM	139	MF801403.1.MEX
78	KY785410.1.BRA	109	KY785455.1.BRA	140	MF801404.1.MEX
79	KY785412.1.USA	110	KY785456.1.BRA	141	MF801406.1.MEX
80	KY785413.1.DOM	111	KY785457.1.USA	142	MF801407.1.MEX
81	KY785414.1.HND	112	KY785459.1.USA	143	MF801408.1.MEX
82	KY785415.1.DOM	113	KY785461.1.HND	144	MF801409.1.MEX
83	KY785417.1.COL	114	KY785462.1.PRI	145	MF801410.1.MEX
84	KY785418.1.HND	115	KY785463.1.DOM	146	MF801412.1.MEX
85	KY785419.1.JAM	116	KY785464.1.PRI	147	MF801413.1.MEX
86	KY785420.1.DOM	117	KY785465.1.DOM	148	MF801414.1.MEX
87	KY785423.1.DOM	118	KY785466.1.COL	149	MF801417.1.MEX
88	KY785424.1.JAM	119	KY785469.1.COL	150	MF801418.1.MEX
89	KY785426.1.BRA	120	KY785470.1.DOM	151	MF801420.1.MEX
90	KY785427.1.BRA	121	KY785474.1.USA	152	MF801423.1.MEX
91	KY785429.1.BRA	122	KY785475.1.DOM	153	MF801426.1.NIC
92	KY785430.1.JAM	123	KY785476.1.DOM		
93	KY785432.1.JAM	124	KY785479.1.BRA		
94	KY785433.1.BRA	125	KY785484.1.DOM		

Tabela B.11 – 1 de 2 - Análise descritiva das partes do modelo de partições estimado para as 153 sequências genômicas completas de vírus da Zika. Da esquerda para a direita, as colunas representam: (1) s: estado s, (2) L_i (# seq): a quantidade entre parêntesis é o número de sequências que fornecem o estado s para a parte L_i. **Nota**: A soma dos números entre parêntesis, por linha, é igual a 153.

s	L_i (# seq)
aaa	L_1 (148), L_2 (2), L_3 (2), L_4 (1).
aac	L_5 (137), L_7 (6), L_8 (4), L_6 (3), L_9 (2), L_{10} (1).
aag	L_{11} (121), L_{13} (27), L_{12} (5).
aat	L_{14} (88), L_{16} (31), L_{17} (29), L_{15} (2), L_{18} (2), L_{19} (1).
aca	L_{20} (138), L_{22} (4), L_{24} (3), L_{21} (2), L_{23} (2), L_{25} (2), L_{26} (2).
acc	L_8 (136), L_5 (9), L_9 (4), L_7 (3), L_6 (1).
acg	L_{27} (119), L_{20} (12), L_{18} (4), L_{28} (4), L_{22} (3), L_{12} (2), L_{29} (2), L_{30} (2), L_{13} (1), L_{24} (1), L_{31} (1), L_{32} (1), L_{33} (1).
act	L_{34} (134), L_{40} (6), L_{38} (4), L_{36} (3), L_{39} (3), L_{37} (2), L_{35} (1).
aga	L_{21} (145), L_{26} (5), L_{19} (1), L_2 (1), L_{31} (1).
agc	L_{41} (129), L_{43} (8), L_6 (7), L_{42} (5), L_{45} (2), L_{44} (1), L_5 (1).
agg	L_{10} (135), L_{29} (15), L_{12} (1), L_{26} (1), L_{31} (1).
agt	L_{38} (120), L_{37} (18), L_{46} (7), L_{36} (2), L_{39} (2), L_{40} (2), L_{34} (1), L_{47} (1).
ata	L_{30} (109), L_{23} (16), L_{27} (12), L_{33} (3), L_{18} (2), L_{26} (2), L_{28} (2), L_{48} (2), L_{17} (1), L_{19} (1), L_{20} (1), L_{21} (1), L_{35} (1).
atc	L_{49} (152), L_{42} (1).
atg	L_{16} (140), L_{17} (12), L_{14} (1).
att	L_{38} (132), L_{34} (7), L_{37} (5), L_{36} (4), L_{40} (3), L_{17} (1), L_{46} (1).
caa	L_{18} (138), L_{26} (5), L_{27} (5), L_{20} (3), L_{30} (1), L_{36} (1).
cac	L_9 (139), L_{44} (5), L_5 (4), L_6 (3), L_{50} (1), L_8 (1).
cag	L_{28} (146), L_{22} (3), L_{20} (2), L_{24} (1), L_{32} (1).
cat	L_{36} (142), L_{19} (2), L_{30} (2), L_{40} (2), L_{17} (1), L_{21} (1), L_{27} (1), L_{35} (1), L_{38} (1).
cca	L_{22} (141), L_{35} (6), L_{20} (2), L_{25} (2), L_{24} (1), L_{51} (1).
ccc	L_6 (137), L_5 (4), L_8 (4), L_{43} (2), L_9 (2), L_{44} (1), L_{45} (1), L_{51} (1), L_{52} (1).
ccg	L_{32} (138), L_{48} (9), L_{33} (3), L_{23} (1), L_{40} (1), L_{51} (1).
cct	L_{46} (138), L_{37} (7), L_{55} (3), L_{38} (2), L_{34} (1), L_{53} (1), L_{54} (1).
cga	L_{15} (132), L_{17} (6), L_{33} (6), L_{47} (4), L_{14} (3), L_{30} (1), L_{39} (1).
cgc	L_{45} (137), L_{25} (6), L_{10} (4), L_{24} (2), L_{44} (1), L_{51} (1), L_6 (1), L_9 (1).
cgg	L_{12} (145), L_{11} (3), L_{10} (2), L_8 (2), L_{13} (1).
cgt	L_{38} (118), L_{46} (10), L_{55} (9), L_{17} (2), L_{19} (2), L_{36} (2), L_{54} (2), L_{15} (1), L_{27} (1), L_{33} (1), L_{34} (1), L_{37} (1), L_{47} (1), L_{48} (1), L_{53} (1).
cta	L_{23} (84), L_{36} (50), L_{35} (6), L_{22} (4), L_{27} (2), L_{21} (1), L_{30} (1), L_{32} (1), L_{38} (1), L_{46} (1), L_{48} (1), L_{51} (1).
ctc	L_{43} (129), L_9 (10), L_{50} (8), L_{42} (2), L_{41} (1), L_{49} (1), L_{52} (1), L_8 (1).
ctg	L_{17} (143), L_{14} (4), L_{16} (3), L_{55} (2), L_{19} (1).
ctt	L_{37} (129), L_{46} (15), L_{38} (2), L_{55} (2), L_{16} (1), L_{34} (1), L_{36} (1), L_{47} (1), L_{53} (1).

Tabela B.12 – 2 de 2 - Análise descritiva das partes do modelo de partições estimado para as 153 sequências genômicas completas de VZ. Da esquerda para a direita: (1) s: estado s, (2) L_i (# seq): entre parêntesis é o número de sequências que fornecem o estado s para a parte L_i. **Nota**: A soma dos números entre parêntesis, por linha, é igual a 153.

s	L_i (# seq)
gaa	L_4 (145), L_1 (2), L_{16} (2), L_2 (2), L_{26} (1), L_3 (1).
gac	L_9 (140), L_5 (9), L_6 (3), L_{44} (1).
gag	L_{31} (146), L_{10} (3), L_2 (3), L_{25} (1).
gat	L_{47} (143), L_{33} (3), L_{53} (3), L_{38} (2), L_{39} (1), L_{55} (1).
gca	L_{26} (143), L_{18} (3), L_{19} (2), L_3 (2), L_{20} (1), L_{28} (1), L_4 (1).
gcc	L_5 (119), L_8 (18), L_9 (10), L_6 (3), L_7 (2), L_{44} (1).
gcg	L_{15} (130), L_{33} (18), L_{38} (2), L_{17} (1), L_{27} (1), L_{47} (1).
gct	L_{55} (143), L_{17} (4), L_{14} (2), L_{54} (2), L_{47} (1), L_{53} (1).
gga	L_2 (147), L_{26} (2), L_3 (2), L_{13} (1), L_{21} (1).
ggc	L_{42} (146), L_{43} (3), L_{41} (2), L_{50} (1), L_{52} (1).
ggg	L_{13} (133), L_{11} (8), L_{12} (8), L_{29} (2), L_2 (1), L_6 (1).
ggt	L_{40} (138), L_{34} (9), L_{38} (3), L_{36} (2), L_{37} (1).
gta	L_{51} (141), L_{32} (7), L_{22} (2), L_{52} (2), L_{42} (1).
gtc	L_7 (145), L_8 (5), L_9 (2), L_5 (1).
gtg	L_{19} (141), L_{17} (9), L_{16} (1), L_{46} (1), L_{54} (1).
gtt	L_{33} (141), L_{15} (7), L_{47} (5).
taa	L_{35} (129), L_{25} (10), L_{22} (7), L_{20} (1), L_{21} (1), L_{41} (1), L_{42} (1), L_{44} (1), L_{48} (1), L_{52} (1).
tac	L_{50} (120), L_{43} (18), L_{44} (4), L_9 (4), L_{42} (2), L_6 (2), L_{41} (1), L_{45} (1), L_8 (1).
tag	L_3 (129), L_1 (5), L_{11} (4), L_4 (4), L_2 (3), L_{13} (2), L_{18} (2), L_{17} (1), L_{26} (1), L_{27} (1), L_{46} (1).
tat	L_{54} (142), L_{21} (4), L_{37} (2), L_{46} (2), L_{16} (1), L_{17} (1), L_{55} (1).
tca	L_{25} (146), L_{35} (3), L_{45} (3), L_6 (1).
tcc	L_{44} (120), L_9 (19), L_{45} (3), L_{41} (2), L_{43} (2), L_{50} (2), L_6 (2), L_{49} (1), L_5 (1), L_7 (1).
tcg	L_{35} (130), L_{22} (13), L_{25} (2), L_{26} (1), L_{28} (1), L_{30} (1), L_{36} (1), L_{40} (1), L_{51} (1), L_{52} (1), L_6 (1).
tct	L_{23} (121), L_{36} (13), L_{30} (10), L_{48} (4), L_{22} (2), L_{21} (1), L_{25} (1), L_{27} (1).
tga	L_{24} (141), L_{45} (5), L_{28} (3), L_{10} (1), L_{20} (1), L_{29} (1), L_6 (1).
tgc	L_{52} (146), L_{42} (7).
tgg	L_{29} (143), L_{10} (9), L_{31} (1).
tgt	L_{53} (141), L_{47} (8), L_{55} (3), L_{15} (1).
tta	L_{48} (132), L_{32} (5), L_{22} (3), L_{35} (3), L_{33} (2), L_{51} (2), L_{21} (1), L_{23} (1), L_{28} (1), L_{30} (1), L_{36} (1), L_{40} (1).
ttc	L_5 (78), L_9 (60), L_6 (5), L_{44} (4), L_{10} (1), L_{41} (1), L_{45} (1), L_{50} (1), L_7 (1), L_8 (1).
ttg	L_{21} (152), L_4 (1).
ttt	L_{39} (142), L_{37} (8), L_{53} (2), L_{38} (1).

Tabela B.13 – 1 de 2 - Análise descritiva das partes do modelo de partições estimado para as 153 sequências genômicas completas de vírus da Zika. Da esquerda para a direita, as colunas representam: (1) L_i: i-ésima parte do modelo de partições, (2) s (freq): a quantidade entre parêntesis é o número de elementos de L_i que possuem o estado s em sua composição. **Nota**: A soma dos números entre parêntesis, por linha, é igual a $|L_i|$.

L_i	s (freq)
1	aaa (148), tag (5), gaa (2).
2	gga (147), gag (3), tag (3), aaa (2), gaa (2), aga (1), ggg (1).
3	tag (129), aaa (2), gca (2), gga (2), gaa (1).
4	gaa (145), tag (4), aaa (1), gca (1), ttg (1).
5	aac (137), gcc (119), ttc (78), acc (9), gac (9), cac (4), ccc (4), agc (1), gtc (1), tcc (1).
6	ccc (137), agc (7), ttc (5), aac (3), cac (3), gac (3), gcc (3), tac (2), tcc (2), acc (1), cgc (1), ggg (1), tca (1), tcg (1), tga (1).
7	gtc (145), aac (6), acc (3), gcc (2), tcc (1), ttc (1).
8	acc (136), gcc (18), gtc (5), aac (4), ccc (4), cgg (2), cac (1), ctc (1), tac (1), ttc (1).
9	gac (140), cac (139), ttc (60), tcc (19), ctc (10), gcc (10), acc (4), tac (4), aac (2), ccc (2), gtc (2), cgc (1).
10	agg (135), tgg (9), cgc (4), gag (3), cgg (2), aac (1), tga (1), ttc (1).
11	aag (121), ggg (8), tag (4), cgg (3).
12	cgg (145), ggg (8), aag (5), acg (2), agg (1).
13	ggg (133), aag (27), tag (2), acg (1), cgg (1), gga (1).
14	aat (88), ctg (4), cga (3), gct (2), atg (1).
15	cga (132), gcg (130), gtt (7), aat (2), cgt (1), tgt (1).
16	atg (140), aat (31), ctg (3), gaa (2), ctt (1), gtg (1), tat (1).
17	ctg (143), aat (29), atg (12), gtg (9), cga (6), gct (4), cgt (2), ata (1), att (1), cat (1), gcg (1), tag (1), tat (1).
18	caa (138), acg (4), gca (3), aat (2), ata (2), tag (2).
19	gtg (141), cat (2), cgt (2), gca (2), aat (1), aga (1), ata (1), ctg (1).
20	aca (138), acg (12), caa (3), cag (2), cca (2), ata (1), gca (1), taa (1), tga (1).
21	ttg (152), aga (145), tat (4), aca (2), ata (1), cat (1), cta (1), gga (1), taa (1), tct (1), tta (1).
22	cca (141), tcg (13), taa (7), aca (4), cta (4), acg (3), cag (3), tta (3), gta (2), tct (2).
23	tct (121), cta (84), ata (16), aca (2), ccg (1), tta (1).
24	tga (141), aca (3), cgc (2), acg (1), cag (1), cca (1).
25	tca (146), taa (10), cgc (6), aca (2), cca (2), tcg (2), gag (1), tct (1).
26	gca (143), aga (5), caa (5), aca (2), ata (2), gga (2), agg (1), gaa (1), tag (1), tcg (1).
27	acg (119), ata (12), caa (5), cta (2), cat (1), cgt (1), gcg (1), tag (1), tct (1).
28	cag (146), acg (4), tga (3), ata (2), gca (1), tcg (1), tta (1).

Tabela B.14 – 2 de 2 - Análise descritiva das partes do modelo de partições estimado para as 153 sequências genômicas completas de vírus da Zika. Da esquerda para a direita, as colunas representam: (1) L_i: i-ésima parte do modelo de partições, (2) s (freq): a quantidade entre parêntesis é o número de elementos de L_i que possuem o estado s em sua composição. **Nota**: A soma dos números entre parêntesis, por linha, é igual a $|L_i|$.

L_i	s (freq)
29	tgg (143), agg (15), acg (2), ggg (2), tga (1).
30	ata (109), tct (10), acg (2), cat (2), caa (1), cga (1), cta (1), tcg (1), tta (1).
31	gag (146), acg (1), aga (1), agg (1), tgg (1).
32	ccg (138), gta (7), tta (5), acg (1), cag (1), cta (1).
33	gtt (141), gcg (18), cga (6), ata (3), ccg (3), gat (3), tta (2), acg (1), cgt (1).
34	act (134), ggt (9), att (7), agt (1), cct (1), cgt (1), ctt (1).
35	tcg (130), taa (129), cca (6), cta (6), tca (3), tta (3), act (1), ata (1), cat (1).
36	cat (142), cta (50), tct (13), att (4), act (3), agt (2), cgt (2), ggt (2), caa (1), ctt (1), tcg (1), tta (1).
37	ctt (129), agt (18), ttt (8), cct (7), att (5), act (2), tat (2), cgt (1), ggt (1).
38	att (132), agt (120), cgt (118), act (4), ggt (3), cct (2), ctt (2), gat (2), gcg (2), cat (1), cta (1), ttt (1).
39	ttt (142), act (3), agt (2), cga (1), gat (1).
40	ggt (138), act (6), att (3), agt (2), cat (2), ccg (1), tcg (1), tta (1).
41	agc (129), ggc (2), tcc (2), ctc (1), taa (1), tac (1), ttc (1).
42	ggc (146), tgc (7), agc (5), ctc (2), tac (2), atc (1), gta (1), taa (1).
43	ctc (129), tac (18), agc (8), ggc (3), ccc (2), tcc (2).
44	tcc (120), cac (5), tac (4), ttc (4), agc (1), ccc (1), cgc (1), gac (1), gcc (1), taa (1).
45	cgc (137), tga (5), tca (3), tcc (3), agc (2), ccc (1), tac (1), ttc (1).
46	cct (138), ctt (15), cgt (10), agt (7), tat (2), att (1), cta (1), gtg (1), tag (1).
47	gat (143), tgt (8), gtt (5), cga (4), agt (1), cgt (1), ctt (1), gcg (1), gct (1).
48	tta (132), ccg (9), tct (4), ata (2), cgt (1), cta (1), taa (1).
49	atc (152), ctc (1), tcc (1).
50	tac (120), ctc (8), tcc (2), cac (1), ggc (1), ttc (1).
51	gta (141), tta (2), cca (1), ccc (1), ccg (1), cgc (1), cta (1), tcg (1).
52	tgc (146), gta (2), ccc (1), ctc (1), ggc (1), taa (1), tcg (1).
53	tgt (141), gat (3), ttt (2), cct (1), cgt (1), ctt (1), gct (1).
54	tat (142), cgt (2), gct (2), cct (1), gtg (1).
55	gct (143), cgt (9), cct (3), tgt (3), ctg (2), ctt (2), gat (1), tat (1).

APÊNDICE C – Obtenção da partição markoviana de M através de $d((i, s),(j, r))$

Para encontrar as partes de M, usando a métrica $d((i,s),(j,r)$ da Definição 8 e o Teorema 9, um método de clusterização hierárquico pode ser utilizado. Os passos do método usado nesta tese são descritos adiante:

1. Inicialmente considera-se cada elemento m_j de M, $j \in \{1,2,...,|M|\}$ como sendo um *cluster*, ou seja, $M = \{m_1, m_2, ..., m_{|M|}\}$ onde $M = \{1,...,p\} \times A^o$. Com base nesses elementos, constrói-se uma matriz de distâncias $D = \{d((i,s),(j,r))\}$ onde $(i,s),(j,r) \in M$ cujas distâncias são dadas pela Definição 8.

2. Buscamos na matriz D pelo par mais similiar $(i_\star, j_\star)$, ou seja,

$$(i_\star, j_\star) = \arg\min\{d(m_i, m_j),\ i \neq j,\ i,j \in \{1,2,...,|M|\}\}$$

   sendo d dado pela Definição 8.

3. Os elementos $i_\star$ e $j_\star$ são juntados num mesmo *cluster*. Assim, a nova configuração de $\mathcal{L}$ é dada por:

$$\mathcal{L}^\star = [[M \setminus \{m_{i_\star}\}] \setminus \{m_{j_\star}\}] \cup m_{i_\star j_\star}$$

   onde $m_{i_\star j_\star} = \{m_{i_\star}, m_{j_\star}\}$. Atualize a matriz de distâncias D:

   (a) Apagando as linhas e colunas correspondentes aos elementos $i_\star$ e $j_\star$, respectivamente.

   (b) Adicionando uma linha e coluna dando as distâncias entre m_k e $m_{i_\star j_\star}$ onde $k \in \{1,2,...,|M|-1\}$, $k \neq i_\star, j_\star$. As distâncias entre m_k e $m_{i_\star j_\star}$ são dadas de acordo com o seguinte método aglomerativo (ou função de ligação) [1]:

$$d(m_k, m_{i_\star j_\star}) = \frac{2}{(|A|-1)\ln(\sum_{l=1}^{p} n_l)} \sum_{a \in A} \left\{ N(m_k,a) \ln \frac{N(m_k,a)}{N(m_k)} + N(m_{i_\star j_\star},a) \ln \frac{N(m_{i_\star j_\star},a)}{N(m_{i_\star j_\star})} - N(\{(m_k, m_{i_\star j_\star})\},a) \ln \frac{N(\{(m_k, m_{i_\star j_\star})\},a)}{N(\{(m_k, m_{i_\star j_\star})\})} \right\} \tag{C.1}$$

   onde $N(\{m_k, m_{i_\star j_\star}\}) = N(m_k) + N(m_{i_\star j_\star})$, $N(\{m_k, m_{i_\star j_\star}\},a) = N(m_k,a) + N(m_{i_\star j_\star},a)$, $N(\{m_{i_\star j_\star}\}) = N(\{m_{i_\star}\}) + N(\{m_{j_\star}\})$ e $N(\{m_{i_\star j_\star}\},a) = N(\{m_{i_\star}\},a) + N(\{m_{j_\star}\},a)$.

4. Repita os passos (2) e (3) até que todos os elementos de M sejam agrupados. Registre as identidades dos *clusters* que são unidos e os níveis (distâncias) em que as fusões acontecem.

[1] Isso é possível pois, como estamos assumindo que $m_{i_\star}$ e $m_{j_\star}$ estão numa mesma parte da partição e, portanto, compartilham as mesmas probabilidades de transição de um estado qualquer para um determinado elemento do alfabeto A.

5. Represente os elementos e os níveis (distâncias) em que as fusões ocorrem em uma tabela ou, se possível, num dendrograma. Por fim, utilize como ponto de corte dos *clusters* aqueles cujo $d^\star$ é no máximo 1. Estes farão a vez das partes de um conjunto de partição markoviana de M.

Exemplo 8. *Considere a mesma situação do Exemplo 7. Usando o mesmo procedimento exposto em (C) o conjunto M foi particionado. Inicialmente, consideramos cada elemento* $m_i \in M$ *como sendo uma parte e, com base na função de ligação proposta (C.1) e, estes são unidos até a formação de um único grupo.*

As Tabelas C.1, C.2, ..., C.11 contém as matrizes contendo as distâncias entre os elementos $m_i \in M$ *atualizados usando (C.1).*

Tabela C.1 – Idêntica a Tabela 32.

	(1,00)	(1,01)	(1,10)	(1,11)	(2,00)	(2,01)
(1,01)	1,47705	-	-	-	-	-
(1,10)	4,39796	1,45304	-	-	-	-
(1,11)	1,45754	0,02676	2,41919	-	-	-
(2,00)	0,91938	0,07128	1,89825	0,02518	-	-
(2,01)	0,10270	1,49591	5,71185	1,62675	0,74125	-
(2,10)	0,00173	2,58294	7,63857	2,98536	1,47556	0,14617
(2,11)	2,33543	0,10737	1,10703	0,37257	0,35356	2,99246
(3,00)	0,00448	0,62211	2,05959	0,55477	0,38806	0,02341
(3,01)	1,20500	0,00910	1,49718	**0,00157**	0,02652	1,09522
(3,10)	6,29312	3,12858	0,45326	4,50643	3,61113	8,03448
(3,11)	4,66197	1,48786	0,07275	3,20672	1,96753	7,28768

	(2,10)	(2,11)	(3,00)	(3,01)	(3,10)
(1,01)	-	-	-	-	-
(1,10)	-	-	-	-	-
(1,11)	-	-	-	-	-
(2,00)	-	-	-	-	-
(2,01)	-	-	-	-	-
(2,10)	-	-	-	-	-
(2,11)	4,71468	-	-	-	-
(3,00)	0,00202	0,94345	-	-	-
(3,01)	1,97384	0,16583	0,51985	-	-
(3,10)	10,06532	2,80936	3,22627	3,11145	-
(3,11)	10,03429	1,16343	1,92368	1,48918	1,03576

As candidatas a partições formadas, em cada uma das iterações, juntamente com o valor de d que houve a junção (que é o menor valor positivo obtido) e o BIC da partição, são descritas na Tabela C.12.

Tabela C.2 – Valores $d((i,s),(j,r))$ (ver Definição 8) calculados entre os elementos (i,s) e (j,r), onde $(i,s),(j,r) \in \bar{\mathcal{P}}(M)$, usando $x_{1,1}^{1001}$, $x_{2,1}^{1002}$ e $x_{3,1}^{1003}$ dadas pelas Tabelas B.1, B.2 e B.3, contidas no anexo, respectivamente. Os elementos de $\bar{\mathcal{P}}(M)$ considerados são: (1,00), (1,01), (1,10), (2,00), (2,01), (2,10), (2,11), (3,00), (3,10), (3,11) e $\{(1,11),(3,01)\}$. A distância mínima não nula é dada entre $(1,00)$ e $(2,10)$ (igual a 0,00173), ou seja, o grupo formado nesta iteração é o $\{(1,00),(2,10)\}$. $\bar{\mathcal{P}}(M)$ é o conjunto de todos os subconjuntos de M, ou seja, é o conjunto de partes (ou conjunto potência) de M.

	(1,00)	(1,01)	(1,10)	(2,00)	(2,01)	(2,10)
(1,01)	1.47705	-	-	-	-	-
(1,10)	4.39796	1.45304	-	-	-	-
(2,00)	0.91938	0.07128	1.89825	-	-	-
(2,01)	0.10270	1.49591	5.71185	0.74125	-	-
(2,10)	**0.00173**	2.58294	7.63857	1.47556	0.14617	-
(2,11)	2.33543	0.10737	1.10703	0.35356	2.99246	4.71468
(3,00)	0.00448	0.62211	2.05959	0.38806	0.02341	0.00202
(3,10)	6.29312	3.12858	0.45326	3.61113	8.03448	10.06532
(3,11)	4.66197	1.48786	0.07275	1.96753	7.28768	10.03429
(1,11)-(3,01)	1.53151	0.02524	2.57087	0.03037	1.79788	3.28603

	(2,11)	(3,00)	(3,10)	(3,11)
(1,01)	-	-	-	-
(1,10)	-	-	-	-
(2,00)	-	-	-	-
(2,01)	-	-	-	-
(2,10)	-	-	-	-
(2,11)	-	-	-	-
(3,00)	0.94345	-	-	-
(3,10)	2.80936	3.22627	-	-
(3,11)	1.16343	1.92368	1.03576	-
(1,11)-(3,01)	0.40355	0.57211	4.72548	3.65523

Tabela C.3 – Valores $d((i,s),(j,r))$ (ver Definição 8) calculados entre os elementos (i,s) e (j,r), onde $(i,s),(j,r) \in \bar{\mathcal{P}}(M)$, usando $x_{1,1}^{1001}$, $x_{2,1}^{1002}$ e $x_{3,1}^{1003}$ dadas pelas Tabelas B.1, B.2 e B.3, contidas no anexo, respectivamente. Os elementos de $\bar{\mathcal{P}}(M)$ considerados são: (1,01), (1,10), (2,00), (2,01), (2,10), (2,11), (3,00), (3,10), (3,11), $\{(1,11),(3,01)\}$ e $\{(1,00),(2,10)\}$. A distância mínima não nula é dada entre $(3,00)$ e $\{(1,00),(2,10)\}$ (igual a 0,00289), ou seja, o grupo formado nesta iteração é o $\{(3,00),(1,00),(2,10)\}$.

	(1,01)	(1,10)	(2,00)	(2,01)	(2,11)	(3,00)
(1,10)	1.45304	-	-	-	-	-
(2,00)	0.07128	1.89825	-	-	-	-
(2,01)	1.49591	5.71185	0.74125	-	-	-
(2,11)	0.10737	1.10703	0.35356	2.99246	-	-
(3,00)	0.62211	2.05959	0.38806	0.02341	0.94345	-
(3,10)	3.12858	0.45326	3.61113	8.03448	2.80936	3.22627
(3,11)	1.48786	0.07275	1.96753	7.28768	1.16343	1.92368
(1,11)-(3,01)	0.02523	2.57087	0.03037	1.79788	0.40355	0.57211
(1,00)-(2,10)	3.03561	8.81786	1.70277	0.18219	5.79597	**0.00289**

	(3,10)	(3,11)	(1,11)-(3,01)
(1,10)	-	-	-
(2,00)	-	-	-
(2,01)	-	-	-
(2,11)	-	-	-
(3,00)	-	-	-
(3,10)	-	-	-
(3,11)	1.03576	-	-
(1,11)-(3,01)	4.72548	3.65523	-
(1,00)-(2,10)	11.30550	12.53098	4.18722

Tabela C.4 – Valores $d((i,s),(j,r))$ (ver Definição 8) calculados entre os elementos (i,s) e (j,r), onde $(i,s),(j,r) \in \bar{\mathcal{P}}(M)$, usando $x_{1,1}^{1001}$, $x_{2,1}^{1002}$ e $x_{3,1}^{1003}$ dadas pelas Tabelas B.1, B.2 e B.3, contidas no anexo, respectivamente. Os elementos de $\bar{\mathcal{P}}(M)$ considerados são: (1,01), (1,10), (2,00), (2,01), (2,10), (2,11), (3,10), (3,11), $\{(1,11),(3,01)\}$ e $\{(3,00),(1,00),(2,10)\}$. A distância mínima não nula é dada entre $(1,01)$ e $\{(1,11),(3,01)\}$ (igual a 0,02524), ou seja, o grupo formado nesta iteração é o $\{(1,01),(1,11),(3,01)\}$.

	(1,01)	(1,10)	(2,00)	(2,01)	(2,11)	(3,10)
(1,10)	1.45304	-	-	-	-	-
(2,00)	0.07128	1.89825	-	-	-	-
(2,01)	1.49591	5.71185	0.74125	-	-	-
(2,11)	0.10737	1.10703	0.35356	2.99246	-	-
(3,10)	3.12858	0.45326	3.61113	8.03448	2.80936	-
(3,11)	1.48786	0.07275	1.96753	7.28768	1.16343	1.03577
(1,11)-(3,01)	**0.02524**	2.57087	0.03037	1.79788	0.40355	4.72548
(3,00)-(1,00)-(2,10)	3.13213	9.12082	1.73887	0.18057	6.09330	11.61647

	(3,11)	(1,11)-(3,01)
(1,10)	-	-
(2,00)	-	-
(2,01)	-	-
(2,11)	-	-
(3,10)	-	-
(3,11)	-	-
(1,11)-(3,01)	3.65523	-
(3,00)-(1,00)-(2,10)	13.31255	4.43951

Tabela C.5 – Valores $d((i,s),(j,r))$ (ver Definição 8) calculados entre os elementos (i,s) e (j,r), onde $(i,s),(j,r) \in \bar{\mathcal{P}}(M)$, usando $x_{1,1}^{1001}$, $x_{2,1}^{1002}$ e $x_{3,1}^{1003}$ dadas pelas Tabelas B.1, B.2 e B.3, contidas no anexo, respectivamente. Os elementos de $\bar{\mathcal{P}}(M)$ considerados são: (1,10), (2,00), (2,01), (2,10), (2,11), (3,10), (3,11), $\{(3,00),(1,00),(2,10)\}$ e $\{(1,01),(1,11),(3,01)\}$. A distância mínima não nula é dada entre $(2,00)$ e $\{(1,01),(1,11),(3,01)\}$ (igual a 0,04613), ou seja, o grupo formado nesta iteração é o $\{(2,00),(1,01),(1,11),(3,01)\}$.

	(1,10)	(2,00)	(2,01)	(2,11)	(3,10)	(3,11)
(2,00)	1.89825	-	-	-	-	-
(2,01)	5.71185	0.74125	-	-	-	-
(2,11)	1.10703	0.35356	2.99246	-	-	-
(3,10)	0.45326	3.61113	8.03448	2.80936	-	-
(3,11)	0.07275	1.96753	7.28768	1.16343	1.03576	-
(3,00)-(1,00)-(2,10)	9.12082	1.73887	0.18057	6.09330	11.61647	13.31255
(1,01)-(1,11)-(3,01)	2.59719	**0.04613**	2.03266	0.37992	4.78594	3.90042

	(3,00)-(1,00)-(2,10)
(2,00)	-
(2,01)	-
(2,11)	-
(3,10)	-
(3,11)	-
(3,00)-(1,00)-(2,10)	-
(1,01)-(1,11)-(3,01)	5.06099

Tabela C.6 – Valores $d((i,s),(j,r))$ (ver Definição 8) calculados entre os elementos (i,s) e (j,r), onde $(i,s),(j,r) \in \bar{\mathcal{P}}(M)$, usando $x_{1,1}^{1001}$, $x_{2,1}^{1002}$ e $x_{3,1}^{1003}$ dadas pelas Tabelas B.1, B.2 e B.3, contidas no anexo, respectivamente. Os elementos de $\bar{\mathcal{P}}(M)$ considerados são: (1,10), (2,01), (2,10), (2,11), (3,10), (3,11), $\{(3,00),(1,00),(2,10)\}$ e $\{(2,00),(1,01),(1,11),(3,01)\}$. A distância mínima não nula é dada entre $(1,10)$ e $(3,11)$ (igual a 0,07275), ou seja, o grupo formado nesta iteração é o $\{(1,10),(3,11)\}$.

	(1,10)	(2,01)	(2,11)	(3,10)	(3,11)	(3,00)-(1,00)-(2,10)
(2,01)	5.71185	-	-	-	-	-
(2,11)	1.10703	2.99246	-	-	-	-
(3,10)	0.45326	8.03448	2.80936	-	-	-
(3,11)	**0.07275**	7.28768	1.16343	1.03576	-	-
(3,00)-(1,00)-(2,10)	9.12082	0.18057	6.09330	11.61647	13.31255	-
(2,00)-(1,01)-(1,11)-(3,01)	2.79244	1.98798	0.46471	5.02623	4.38169	5.08152

Tabela C.7 – Valores $d((i,s),(j,r))$ (ver Definição 8) calculados entre os elementos (i,s) e (j,r), onde $(i,s),(j,r) \in \bar{\mathcal{P}}(M)$, usando $x_{1,1}^{1001}$, $x_{2,1}^{1002}$ e $x_{3,1}^{1003}$ dadas pelas Tabelas B.1, B.2 e B.3, contidas no anexo, respectivamente. Os elementos de $\bar{\mathcal{P}}(M)$ considerados são: (2,01), (2,11), (3,10), $\{(3,00),(1,00),(2,10)\}$, $\{(2,00),(1,01),(1,11),(3,01)\}$ e $\{(1,10),(3,11)\}$. A distância mínima não nula é dada entre $(2,01)$ e $\{(3,00),(1,00),(2,10)\}$ (igual a 0,18057), ou seja, o grupo formado nesta iteração é o $\{(2,01),(3,00),(1,00),(2,10)\}$.

	(2,01)	(2,11)	(3,10)	(3,00)-(1,00)-(2,10)	(2,00)-(1,01) (1,11)-(3,01)
(2,11)	2,99246	-	-	-	-
(3,10)	8,03448	2,80936	-	-	-
(3,00)-(1,00)-(2,10)	**0,18057**	6,09330	11,61647	-	-
(2,00)-(1,01)-(1,11)-(3,01)	1,98798	0,46471	5,02623	5,08152	-
(1,10)-(3,11)	8,27530	1,52195	0,96584	15,37606	5,78428

Tabela C.8 – Valores $d((i,s),(j,r))$ (ver Definição 8) calculados entre os elementos (i,s) e (j,r), onde $(i,s),(j,r) \in \bar{\mathcal{P}}(M)$, usando $x_{1,1}^{1001}$, $x_{2,1}^{1002}$ e $x_{3,1}^{1003}$ dadas pelas Tabelas B.1, B.2 e B.3, contidas no anexo, respectivamente. Os elementos de $\bar{\mathcal{P}}(M)$ considerados são: (2,11), (3,10), $\{(2,01),(3,00),(1,00),(2,10)\}$, $\{(2,00),(1,01),(1,11),(3,01)\}$ e $\{(1,10),(3,11)\}$. A distância mínima não nula é dada entre $(2,11)$ e $\{(2,00)$, $(1,01)$, $(1,11)$, $(3,01)\}$ (igual a 0,46471), ou seja, o grupo formado nesta iteração é o $\{(2,11)$, $(2,00)$, $(1,01)$, $(1,11)$, $(3,01)\}$.

	(2,11)	(3,10)	(2,00)-(1,01)-(1,11)-(3,01)	(1,10)-(3,11)
(3,10)	2.80936	-	-	-
(2,00)-(1,01)-(1,11)-(3,01)	**0,46471**	5,02623	-	-
(1,10)-(3,11)	1,52195	0,96584	5,78428	-
(2,01)-(3,00)-(1,00)-(2,10)	6,60079	12,06725	5,83201	18,80241

Tabela C.9 – Valores $d((i,s),(j,r))$ (ver Definição 8) calculados entre os elementos (i,s) e (j,r), onde $(i,s),(j,r) \in \bar{\mathcal{P}}(M)$, usando $x_{1,1}^{1001}$, $x_{2,1}^{1002}$ e $x_{3,1}^{1003}$ dadas pelas Tabelas B.1, B.2 e B.3, contidas no anexo, respectivamente. Os elementos de $\bar{\mathcal{P}}(M)$ considerados são: (3,10), $\{(2,01),(3,00),(1,00),(2,10)\}$, $\{(2,11),(2,00),(1,01),(1,11),(3,01)\}$ e $\{(1,10),(3,11)\}$. A distância mínima não nula é dada entre $(3,10)$ e $\{(1,10),(3,11)\}$ (igual a 0,96584), ou seja, o grupo formado nesta iteração é o $\{(3,10),(1,10),(3,11)\}$.

	(3,10)	(1,10)-(3,11)	(2,01)-(3,00)-(1,00)-(2,10)
(1,10)-(3,11)	**0,96584**	-	-
(2,01)-(3,00)-(1,00)-(2,10)	12,06725	18,80241	-
(2,11)-(2,00)-(1,01)-(1,11)-(3,01)	4,64041	5,34572	7,68539

Tabela C.10 – Valores $d((i,s),(j,r))$ (ver Definição 8) calculados entre os elementos (i,s) e (j,r), onde $(i,s),(j,r) \in \bar{\mathcal{P}}(M)$, usando $x_{1,1}^{1001}$, $x_{2,1}^{1002}$ e $x_{3,1}^{1003}$ dadas pelas Tabelas B.1, B.2 e B.3, contidas no anexo, respectivamente. Os elementos de $\bar{\mathcal{P}}(M)$ considerados são: $\{(2,01),(3,00),(1,00),(2,10)\}$, $\{(2,11),(2,00),(1,01),(1,11),(3,01)\}$ e $\{(3,10),(1,10),(3,11)\}$. A distância mínima não nula é dada entre $\{(2,11),(2,00),(1,01),(1,11),(3,01)\}$ e $\{(3,10),(1,10),(3,11)\}$ (igual a 7,57926), ou seja, o grupo formado nesta iteração é o $\{(2,11),(2,00),(1,01),(1,11),(3,01),(3,10),(1,10),(3,11)\}$.

	(2,01)-(3,00)-(1,00)-(2,10)	(2,11)-(2,00)-(1,01)-(1,11)-(3,01)
(2,11)-(2,00)-(1,01)-(1,11)-(3,01)	7,68539	-
(3,10)-(1,10)-(3,11)	22,40396	**7,57926**

Tabela C.11 – Valores $d((i,s),(j,r))$ (ver Definição 8) calculados entre os elementos (i,s) e (j,r), onde $(i,s),(j,r) \in \bar{\mathcal{P}}(M)$, usando $x_{1,1}^{1001}$, $x_{2,1}^{1002}$ e $x_{3,1}^{1003}$ dadas pelas Tabelas B.1, B.2 e B.3, contidas no anexo, respectivamente. Os elementos de $\bar{\mathcal{P}}(M)$ considerados são: $\{(2,01),(3,00),(1,00),(2,10)\}$ e $\{(2,11),(2,00),(1,01),(1,11),(3,01),(3,10),(1,10),(3,11)\}$. A distância mínima não nula é dada entre $\{(2,01),(3,00),(1,00),(2,10)\}$ e $\{(2,11),(2,00),(1,01),(1,11),(3,01),(3,10),(1,10),(3,11)\}$ (igual a 15,07281), ou seja, o grupo formado nesta iteração é o $\{(2,01),(3,00),(1,00),(2,10),(2,11),(2,00),(1,01),(1,11),(3,01),(3,10),(1,10),(3,11)\}$, que é o próprio conjunto M.

	(2,01)-(3,00)-(1,00)-(2,10)
(2,11)-(2,00)-(1,01)-(1,11)-(3,01)-(3,10)-(1,10)-(3,11)	15,07281

Tabela C.12 – Partições formadas através de d usando as amostras $x_{1,1}^{1001}$, $x_{2,1}^{1002}$ e $x_{3,1}^{1003}$ apresentadas nas Tabelas B.1, B.2 e B.3, respectivamente. A partição inicial considerada é aquela onde cada elemento do conjunto M corresponde a exatamente uma parte. A linha negritada corresponde ao maior d_{min} que é inferior a 1. Repare que, essa partição do conjunto M é a que possui o maior BIC dentre as demais.

Configuração das partes nessa etapa	d_{min}	BIC
$\{(1,00)\}, \{(1,01)\}, \{(1,10)\}, \{(2,00)\}, \{(2,01)\}, \{(2,10)\}, \{(2,11)\}, \{(3,00)\}, \{(3,10)\}, \{(3,11)\}, \{(1,11),(3,01)\}$	0,00157	-1741,88
$\{(1,01)\}, \{(1,10)\}, \{(2,00)\}, \{(2,01)\}, \{(2,11)\}, \{(3,00)\}, \{(3,10)\}, \{(3,11)\}, \{(1,11),(3,01)\}, \{(1,00),(2,10)\}$	0,00173	-1737,88
$\{(1,01)\}, \{(1,10)\}, \{(2,00)\}, \{(2,01)\}, \{(2,11)\}, \{(3,10)\}, \{(3,11)\}, \{(1,11),(3,01)\}, \{(3,00),(1,00),(2,10)\}$	0,00289	-1733,89
$\{(1,10)\}, \{(2,00)\}, \{(2,01)\}, \{(2,11)\}, \{(3,10)\}, \{(3,11)\}, \{(3,00),(1,00),(2,10)\}, \{(1,01),(1,11),(3,01)\}$	0,02524	-1729,99
$\{(1,10)\}, \{(2,01)\}, \{(2,11)\}, \{(3,10)\}, \{(3,11)\}, \{(3,00),(1,00),(2,10)\}, \{(2,00),(1,01),(1,11),(3,01)\}$	0,04613	-1726,17
$\{(2,01)\}, \{(2,11)\}, \{(3,10)\}, \{(3,00),(1,00),(2,10)\}, \{(2,00),(1,01),(1,11),(3,01)\}, \{(1,10),(3,11)\}$	0,07275	-1722,46
$\{(2,11)\}, \{(3,10)\}, \{(2,00),(1,01),(1,11),(3,01)\}, \{(1,10),(3,11)\}, \{(2,01),(3,00),(1,00),(2,10)\}$	0,18057	-1719,18
$\{(3,10)\}, \{(1,10),(3,11)\}, \{(2,01),(3,00),(1,00),(2,10)\}, \{(2,11),(2,00),(1,01),(1,11),(3,01)\}$	0,46471	-1717,04
$\{\mathbf{(2,01),(3,00),(1,00),(2,10)}\}, \{\mathbf{(2,11),(2,00),(1,01),(1,11),(3,01)}\}, \{\mathbf{(3,10),(1,10),(3,11)}\}$	**0,96583**	**-1716,90**
$\{(2,01),(3,00),(1,00),(2,10)\}, \{(2,11),(2,00),(1,01),(1,11),(3,01),(3,10),(1,10),(3,11)\}$	7.57926	-1743.25
$\{(2,01),(3,00),(1,00),(2,10),(2,11),(2,00),(1,01),(1,11),(3,01),(3,10),(1,10),(3,11)\}$	15,07281	-1799,60

O dendrograma obtido desse processo de particionamento hierárquico aglomerativo, usando a função de ligação proposta (C.1) é apresentado na Figura C.1. Observamos no dendrograma claramente 3 (três) grupos, a saber: (1) (2,11), (2,00), (3,01), (1,01) e (1,11); (2) (3,00), (2,10), (1,00) e (2,01); (3) (3,10), (1,10) e (3,11). Repare que o terceiro grupo formado está no limiar da distância mínima de 1 e é, justamente, onde o método de particionamento de M proposto errou, ou seja, os elementos (3,10), (1,10) e (3,11) foram unidos numa mesma parte, quando, na verdade, os elementos (1,10) e (3,11) foram uma parte e, (3,11), outra (Vide Tabela 30).

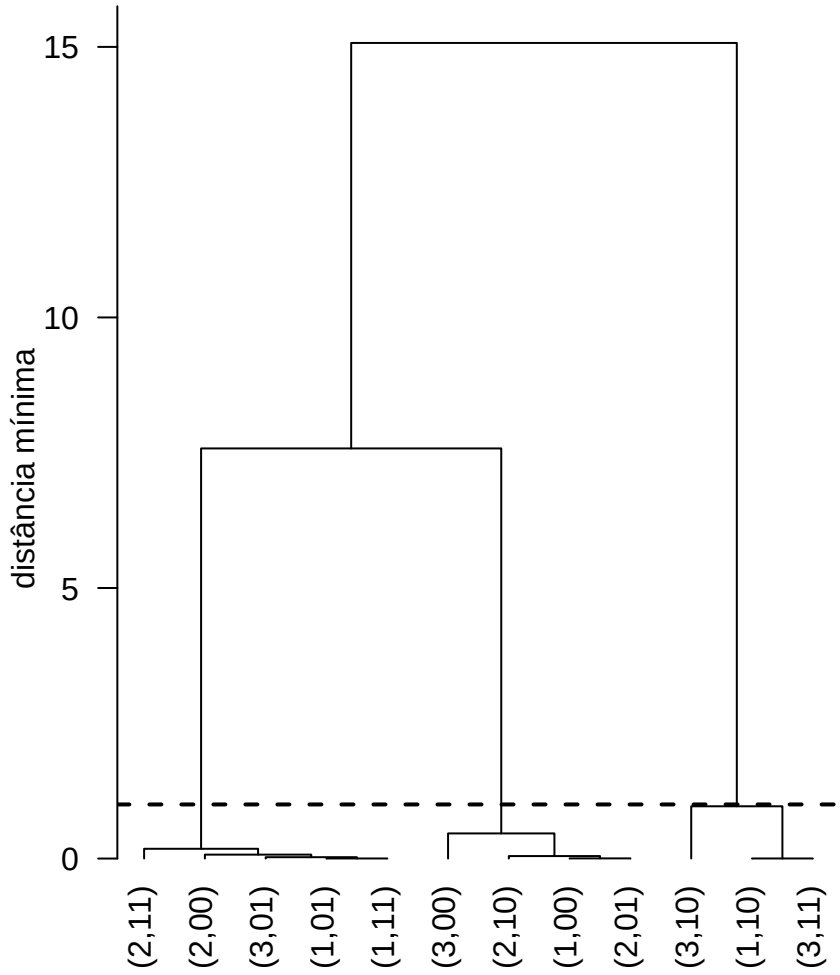


Figura C.1 – Dendrograma obtido do particionamento do conjunto $M = \{1, 2, 3\} \times \{0, 1\}^2$, com a matriz de distâncias construída usando-se $d((i, s), (j, r))$ (ver Definição 8) com a função de ligação proposta na Seção C. As amostras usadas nesse agrupamento são $x_{1,1}^{1001}$, $x_{2,1}^{1002}$ e $x_{3,1}^{1003}$ apresentadas nas Tabelas B.1, B.2 e B.3, respectivamente.

A verdadeira partição de M é apresentada na Tabela 30. Repare que as partes obtidas de M, quando $d^\star = 0,965837$, *coincidem com 3 das quatro partes verdadeiras, onde, as partes* $L_3 = \{(1, 10), (3, 11)\}$ *e* $L_4 = \{(3, 10)\}$ *foram unidas numa única parte, provavelmente pelo fato das probabilidades de transição serem relativamente próximas. Outro fato*

que provavelmente interferiu no resultado obtido é o tamanho relativamente reduzido da amostra, uma vez que, o procedimento de particionamento baseia-se na consistência de $d((i,s),(j,r))$.

Uma outra forma de obter-se uma partição do conjunto M por meio da medida d é através de um método de clusterização hierárquica usando a função de ligação simples, média, completa ou de Ward. Os clusters identificados nos respectivos dendrogramas serão as partições obtidas de M. Dendrogramas aplicado a uma matriz de distâncias cujas entradas são os elementos da Tabela 32, usando as ligações de Ward, simples, média e completa são apresentados na Figura C.2. Os quatro dendrogramas apresentam 3 (três) grupos, não necessariamente nesta ordem, a saber: (1) (3,10), (1,10) e (3,11), (2) (2,11), (2,00), (3,01), (1,01) e (1,11), (3) (3,00), (2,10), (1,00) e (2,01).

Com base na Definição 5, os elementos de uma parte compartilham suas probabilidades. Assim, podemos também estimar suas probabilidades condicionais de transição $P(\cdot|L_i)$ por meio de

$$\hat{P}(\cdot|\hat{L}_i) = \frac{N(\hat{L}_i,\cdot)}{N(\hat{L}_i)} = \frac{\sum_{(j,s)\in\hat{L}_i} N((j,s),\cdot)}{\sum_{(j,s)\in\hat{L}_i} N((j,s))} \quad \text{, onde} \quad \cdot \in A \text{ e } i \in \{1,...,|\mathcal{L}|\}\,. \tag{C.2}$$

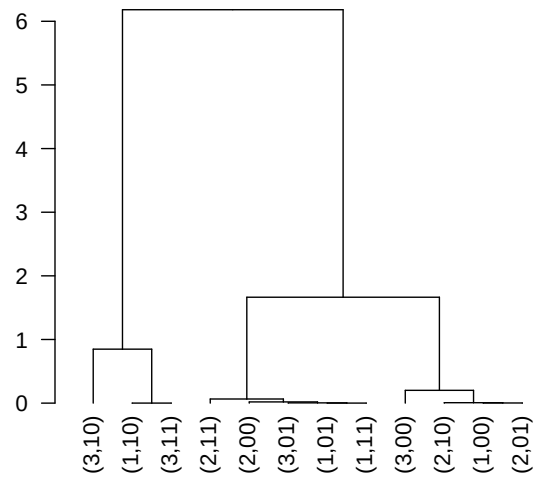
Exemplo 9. *Consideremos a mesma situação dos exemplos 7 e 8. A estimativa* $\hat{\mathcal{L}}$ *da verdadeira partição* $\mathcal{L}$ *de M obtida foi:*

$$\begin{aligned}
&\hat{\mathcal{L}} = \left\{\hat{L}_1, \hat{L}_2, \hat{L}_3\right\} \quad \textit{onde}\\
&\hat{L}_1 = \{(2,01),(3,00),(1,00),(2,10)\}\\
&\hat{L}_2 = \{(2,11),(2,00),(1,01),(1,11),(3,01)\}\\
&\hat{L}_3 = \{(3,10),(1,10),(3,11)\}\,.
\end{aligned}$$

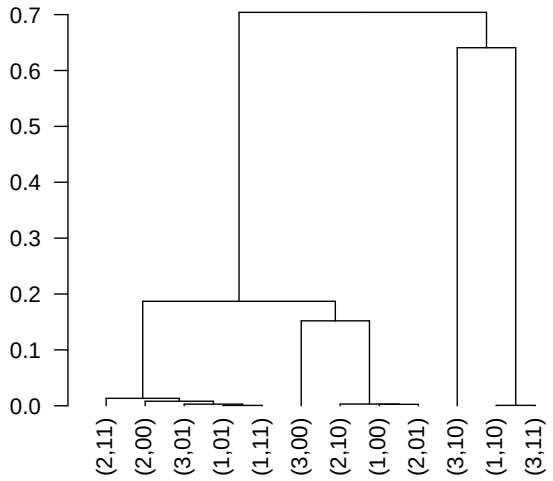
Assim, com base na equação C.2, temos [2]*:*

$$\begin{aligned}
\hat{P}(0|\hat{L}_1) &= \frac{N(\hat{L}_1,0)}{N(\hat{L}_1)} = \frac{\sum_{(j,s)\in\hat{L}_1} N((j,s),\cdot)}{\sum_{(j,s)\in\hat{L}_1} N((j,s))} = \\
&= \frac{N((2,01),0)+N((3,00),0)+N((1,00),0)+N((2,10),0)}{N((2,01))+N((3,00))+N((1,00))+N((2,10))} = \\
&= \frac{95+16+39+106}{95+112+16+16+39+36+106+101} = \frac{256}{521} = 0,49136
\end{aligned}$$

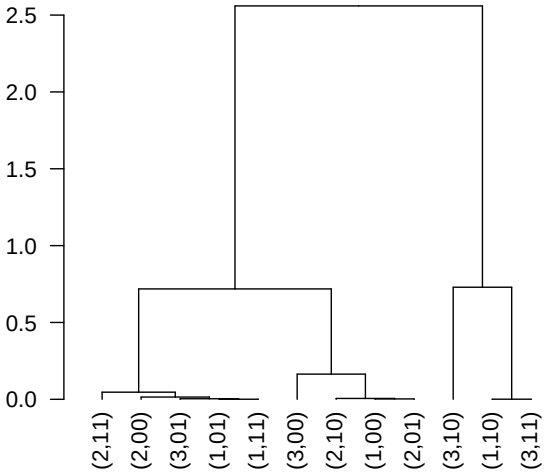
[2] As quantidades $N((2,01),0)$, $N((3,00),0)$, $N((1,00),0)$, etc... utilizadas no Exemplo 9 estão disponíveis na Tabela 31.



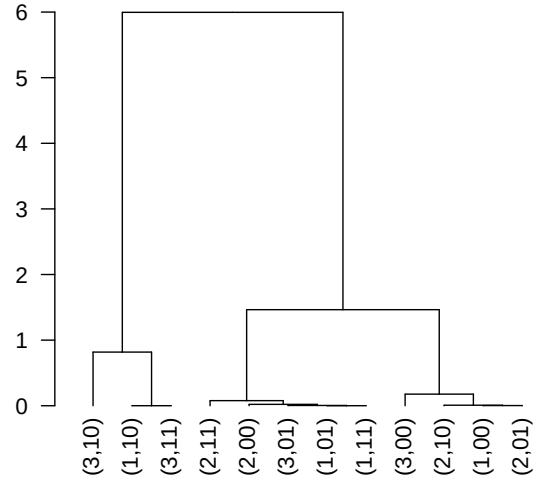
(a) Ligação de Ward.



(a) Ligação Simples.



(a) Ligação Média.



(a) Ligação Completa.

Figura C.2 – Dendrogramas das sequências $x_{1,1}^{1001}$ (Tabela B.1), $x_{2,1}^{1002}$ (Tabela B.2) e $x_{3,1}^{1003}$ (Tabela B.3) construída usando $d((i,s),(j,r))$. A matriz de distâncias é dada na Tabela 32. As funções de ligação utilizadas foram: (a) Ward (*Ward criterion*), (b) Simples, (c) Média (d) Completa.

$$\hat{P}(0|\hat{L}_2) = \frac{N(\hat{L}_2,0)}{N(\hat{L}_2)} = \frac{\sum_{(j,s)\in\hat{L}_2} N((j,s),\cdot)}{\sum_{(j,s)\in\hat{L}_2} N((j,s))} =$$
$$= \frac{N((2,11),0)+N((2,00),0)+N((1,01),0)+N((1,11),0)+N((3,01),0)}{N((2,11))+N((2,00))+N((1,01))+N((1,11))+N((3,01))} =$$
$$= \frac{112+53+65+153+52}{112+315+53+106+65+154+153+334+52+116} = \frac{435}{1460} = 0,29795$$

$$\hat{P}(0|\hat{L}_3) = \frac{N(\hat{L}_3, 0)}{N(\hat{L}_3)} = \frac{\sum_{(j,s)\in\hat{L}_3} N((j,s),\cdot)}{\sum_{(j,s)\in\hat{L}_3} N((j,s))} =$$
$$= \frac{N((3,10),0) + N((1,10),0) + N((3,11),0)}{N((3,10)) + N((1,10)) + N((3,11))} =$$
$$= \frac{16 + 35 + 116}{16 + 152 + 35 + 183 + 116 + 517} = \frac{167}{1090} = 0,15321$$

A Tabela C.13 apresenta as probabilidades condicionais: (a) nas partes estimadas, (b) nas partes verdadeiras, de transição do MMP. Repare que $\hat{L}_1 = L_1$, $\hat{L}_2 = L_2$ *e* $\hat{L}_3 = L_3 \cup L_4$. *Observe também que:*

$$\hat{P}(\cdot|\hat{L}_1) \cong P(\cdot|L_1), \quad \hat{P}(\cdot|\hat{L}_2) \cong P(\cdot|L_2) \quad e \quad \hat{P}(\cdot|\hat{L}_3) \cong 0,5P(\cdot|L_3) + 0,5P(\cdot|L_4).$$

Tabela C.13 – Probabilidades condicionais de transição das partes: (a) estimadas do conjunto $M = \{1,2,3\} \times \{0,1\}^2$ obtidas a partir das três amostras $x_{1,1}^{1001}$ (Tabelas B.1), $x_{2,1}^{1002}$ (Tabelas B.2) e $x_{3,1}^{1003}$ (Tabelas B.3). (b) verdadeiras do conjunto $M = \{1,2,3\} \times \{0,1\}^2$ onde as probabilidade de transição para os processos $(X_{j,t})$, $j = 1,2,3$ são apresentadas na Tabela 29.

Partes estimadas	Elementos	$\hat{P}(0\|\hat{L}_k)$	$\hat{P}(1\|\hat{L}_k)$
$\hat{L}_1$	$(2,01),(3,00),(1,00),(2,10)$	0,491	0,509
$\hat{L}_2$	$(2,11),(2,00),(1,01),(1,11),(3,01)$	0,298	0,702
$\hat{L}_3$	$(3,10),(1,10),(3,11)$	0,153	0,847

(a)

.

Partes verdadeiras	Elementos	$P(0\|L_k)$	$P(1\|L_k)$
L_1	(1,00), (2,01), (2,10), (3,00)	0,50	0,50
L_2	(1,01), (1,11), (2,00), (2,11), (3,01)	0,30	0,70
L_3	(1,10), (3,11)	0,20	0,80
L_4	(3,10)	0,10	0,90

(b)

APÊNDICE D – Artigos publicados

Esta tese deu origem aos seguintes trabalhos:

1. CORDEIRO, M. T. A., GARCÍA, J.E., GONZÁLEZ-LÓPEZ, V.A., LONDOÑO, S.L.M. Partition Markov model for multiple processes. Math Meth Appl Sci. 2020; 43: 7677– 7691. https://doi.org/10.1002/mma.6079
2. CORDEIRO, M. T. A.; GARCÍA, J. E.; GONZÁLEZ-LÓPEZ, V.A.; LONDOÑO, S.L.M.(2019) Stochastic profile of Epstein-Barr virus in nasopharyngeal carcinoma settings, 4open, 2, 25.
3. CORDEIRO, M. T. A.; GARCÍA, J. E.; GONZÁLEZ-LÓPEZ, V.A.; LONDOÑO, S.L.M.(2019) Classification of autochthonous dengue virus type 1 strains circulating in Japan in 2014 4open, 2, 20.
4. SIMILARITY BETWEEN STRAINS OF ZIKA FROM TROPICAL AND SUB-TROPICAL REGIONS. GARCÍA, J. E.; GONZÁLEZ-LÓPEZ, V.A.; LONDOÑO, S.L.M.; CORDEIRO, M.T.A. Data set and applications, ISTE science and publishing LTD. (capítulo de livro aceito para publicação).

Nas páginas a seguir encontram-se dois destes artigos.

D.1 Classification of autochthonous dengue virus type 1 strains circulating in Japan in 2014

4open 2019, **2**, 20

https://doi.org/10.1051/fopen/2019018

4 Open

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

Classification of autochthonous dengue virus type 1 strains circulating in Japan in 2014

Marcos T.A. Cordeiro[1,*], Jesús E. García[2], Verónica A. González-López[2], and Sergio L.M. Londoño[2]

[1] Department of Mathematics, Federal University of Technology, Av. Monteiro Lobato, s/n – Km 04, Campus Ponta Grossa, 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 13 May 2019

Abstract – In this paper, we classify by representativeness the elements of a set of complete genomic sequences of Dengue Virus Type 1 (DENV-1), corresponding to the outbreak in Japan during 2014. The set is coming from four regions: Chiba, Hyogo, Shizuoka and Tokyo. We consider this set as composed of independent samples coming from Markovian processes of finite order and finite alphabet. Under the assumption of the existence of a law that prevails in at least 50% of the samples of the set, we identify the sequences governed by the predominant law (see [1, 2]). The rule of classification is based on a local metric between samples, which tends to zero when we compare sequences of identical law and tends to infinity when comparing sequences with different laws. We found that the order of representativeness, from highest to lowest and according to the origin of the sequences is: Tokyo, Chiba, Hyogo, and Shizuoka. When comparing the Japanese sequences with their contemporaries from Asia, we find that the less representative sequence (from Shizuoka) is positioned in groups considerably far away from that which includes the sequences from the other regions in Japan, this offers evidence to suppose that the outbreak in Japan could be produced by more than one type of DENV-1.

Keywords: Classification of stochastic samples, Metric between stochastic processes

1 Introduction

In the present study, we analyze the entire genome of six autochthonous DENV-1 (Dengue Virus Type 1) strains isolated from patients during the 2014 outbreak in Japan (see [3]). Our objective is to identify the most representative sequence and the least representative sequence of the set. The virus is transmitted to humans by infected mosquitoes of the *Aedes* genus. The Dengue virus, in their four types, can be contracted more than once, what makes it extremely efficient. Individuals who have already contracted Dengue present risk factors that, by contracting the virus for the second time in a different variant/type, can develop severe forms such as *Dengue hemorrhagic fever* and *Dengue shock syndrome* both potentially fatal. In recent years it has been possible to have access to a vaccine, despite this, it is only recommended for those who have already contracted Dengue before. Since the Second World War, the cases of Dengue Fever identified in Japan have been imported, but between 2013 and 2014 more than 160 autochthonous cases were identified. This has demanded a deep investigation of the nature of this outbreak. The contamination has been caused by DENV-1. The findings in [3] suggest that there were at least two independent autochthonous epidemics in Japan in 2014 caused by DENV-1. In this work our objective is to classify the Dengue samples considering the representativeness that each sample has in relation to the group. That is, we know that the samples belong to different individuals and as a consequence are subject to variations in their genomic construction. We wish to identify the most representative sample of the group, this being the most similar to all other samples of the group. Also, we want to identify the most discrepant sample in the group. To achieve our goal, we will identify each sequence with a sample of a stochastic process. Then we will measure the distance between the sequences using a specially designed metric, and applying a robust method (introduced in [1]), we can identify the most representative sample and we can also classify all the samples in order of representativeness. The problem of establishing the proximity between genomic sequences has aroused the interest of several areas and with different objectives. For example, in the area

*Corresponding author: marcoscordeiro@utfpr.edu.br

of virology a relevant issue is the characterization of the Epstein Barr virus under different diagnoses, such as Burkitt's lymphoma, nasopharyngeal carcinoma or even among several types of associated diagnoses [4]. See for examples [5–8], in each case some notion of similarity between strains is explored. An aspect that differentiates this article is that the notion used to establish the proximity between the genomic sequences is a metric (see [2]), that is, it is symmetric, not negative and verifies the triangular inequality. This metric also has very convenient statistical characteristics, such as being statistically consistent. The second aspect and perhaps the most relevant one is the construction of a metric-based classification between the sequences (see [1]). A criterion for classifying samples, statistically consistent as is the case, could be used in the future to construct standard representations for genomic structures. For example, it could certify that the sequence currently used as a gold standard in the Epstein Barr virus study, sequence B95-8, from [9] (see also [5]), in fact, serves impartial criteria such as the one used in this paper.

The sections and topics that compound this article are detailed below. The notions that we use as well as the definition of the criterion to classify the sequences are given in Section 2. We detail the database in Section 3.1. The results are in Section 3.2 and the conclusions in Section 4.

2 Theoretical basis

In this section we give the theoretical framework on which are established, (i) the notion of proximity between sequences as well as (ii) the criterion of classification of the sequences.

Let (X_t) be a discrete time, order o (with $o < \infty$) Markov chain on a finite alphabet A. Let us call $\mathcal{S} = A^o$ the state space, 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 conditional probability $P(a|s) = \text{Prob}\,(X_t = a|X_{t-o}^{t-1} = s)$. In a given sample x_1^n, coming from the stochastic process, the number of occurrences of s in the sample x_1^n is denoted by $N_n(s)$ and the number of occurrences of s followed by a in the sample x_1^n is denoted by $N_n(s, a)$. In this way, $\frac{N_n(s,a)}{N_n(s)}$ is the estimator of $P(a|s)$. Consider now, two Markov chains $(X_{1,t})$ and $(X_{2,t})$, of order o, arranged on the finite alphabet A with state space $\mathcal{S}$. Given $s \in \mathcal{S}$ denote by $\{P(a|s)\}_{a\in A}$ and $\{Q(a|s)\}_{a\in A}$ the sets of conditional probabilities of $(X_{1,t})$ and $(X_{2,t})$ respectively. We define a local metric d_s (introduced by [2]) that, when evaluated in a given string s, allows us to define how far or near the processes are.

Definition 2.1. *Consider two Markov chains $(X_{1,t})$ and $(X_{2,t})$, of order o, with finite alphabet A, state space $\mathcal{S} = A^o$ and independent samples $x_{1,1}^{n_1}$, $x_{2,1}^{n_2}$ respectively.*

(i) *For a string $s \in \mathcal{S}$,*

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

(ii)

$$d_{\max}\,(x_{1,1}^{n_1}, x_{2,1}^{n_2}) = \max_{s\in\mathcal{S}} \{d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2})\},$$

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

The Definition 2.1 offers us two notions of proximity between sequences, "i." is local and "ii." is global, "i." and "ii." are statistically consistent, that is, by increasing the $\min\{n_1, n_2\}$ grows their ability to detect (a) discrepancies (when the underlying laws are different) and (b) similarities (when the underlying laws are the same). In the application we use $\alpha = 2$ (see Definition 2.1-i.), with this value ($\alpha = 2$), to decide that the sequences follow the same law when $d_s < 1$, is equivalent to use the *Bayesian Information Criterion* (see [2, 10]). In [2] is proved also that d_s is a metric:

(a) $d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) \geq 0$ with equality $\iff \frac{N_{n_1}(s,a)}{N_{n_1}(s)} = \frac{N_{n_2}(s,a)}{N_{n_2}(s)}\ \forall a \in A$,

(b) $d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) = d_s(x_{2,1}^{n_2}, x_{1,1}^{n_1})$,

(c) $d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) \leq d_s(x_{1,1}^{n_1}, x_{3,1}^{n_3}) + d_s(x_{3,1}^{n_3}, x_{2,1}^{n_2})$.

To follow is introduced a notion that makes possible the classification of sequences that belong to a group of sequences.

Definition 2.2. *Given a finite collection* $\{x_{j,1}^{n_j}\}_{j=1}^{m}$ *of samples from the processes* $\{(X_{j,t})\}_{j=1}^{m}$ *with probabilities* $\{P_j\}_{j=1}^{m}$, *over the finite alphabet A, with state space* $\mathcal{S} = A^o$ $(o < \infty)$. *For a fixed* $i \in \{1, 2, \ldots, m\}$ *define*

$$V(x_{i,1}^{n_i}) = \textit{median}\,\{d_{max}(x_{i,1}^{n_i}, x_{j,1}^{n_j}) : j \neq i, 1 \leq j \leq m\}.$$

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

With the V values attributed to each sample, we can proceed to order the samples, from lowest to highest value of V, in order to identify their classification. As we can perceive from the Definition 2.2, low values of V indicate that these samples represent the whole group better, while high values of V indicate little representativeness. The next result (proved in [1]), give us an adequate tool to classify sequences, according to their underlying laws, it allows to consolidate V as a robust and consistent classifier.

Theorem 2.1. *Under the assumptions of Definition 2.2, for each i,* $1 \leq i \leq m$, *set* $\xi_i = |\{j : 1 \leq j \leq m, P_j = P_i\}|$,

(i)

$$V\left(x_{i,1}^{n_i}\right) \underset{\min\{n_1,\cdots,n_m\}\to\infty}{\to} \infty, \textit{ if, and only if,} \quad \xi_i \leq \left\lceil \frac{m}{2} \right\rceil.$$

(ii)

$$V(x_{i,1}^{n_i}) \underset{\min\{n_1,\cdots,n_m\}\to\infty}{\to} 0, \textit{ if, and only if,} \quad \xi_i > \left\lceil \frac{m}{2} \right\rceil.$$

where $\lceil x \rceil$ *is the smallest integer greater than or equal to x.* Theorem 2.1 guarantees that if at least 50% of the samples of the set follow the same law, each of them receives a value of V close to zero. And if this does not happen, V takes arbitrarily large values identifying discrepancies in the generating laws of the sequences.

3 Data and results

First, we describe the data, its source and structure, afterwards we proceed to measure the distance between the sequences and to classify them by representativeness.

3.1 Dengue virus type 1

The complete sequences were obtained from http://www.ncbi.nlm.nih.gov/ (NCBI – National Center for Biotechnology Information), sequenced and studied for the first time in [3]. We describe the sequences in Table 1.

The epicenter of Dengue Fever (DF) outbreak during 2014 was possible in the Yoyogi Park in Tokyo. Part of the sequences are coming from patients who pass through there and nearby locations. Details of each patient listed in Table 1 are given in [3], here we emphasize some of them. In the last column of Table 1 we inform the place in where it is suspected that the contamination happened to each patient. The contamination of patient 14-149J occurred in a place near to Yoyogi Park. 14-153J did not visit Yoyogi Park for at least two weeks before the onset of DF and was likely infected in Chiba prefecture. 14-181J lives in Shizuoka prefecture and never visited Yoyogi Park or the other affected areas, visited other places in Tokyo before the onset of DF. Patient 14-188J lives in Nishinomiya city, Hyogo prefecture, over 500 km of Tokyo and never visited the Tokyo area before the onset of DF. He visited Malaysia for seven days and had the onset of DF 12 days after. To illustrate the structure of the data, consider the beginning of the sequence LC011945,

gacaagaacagtttcgaatcggaagcttgcttaacgtagttctaacagtt ...

Table 1. Complete Sequences of Dengue Virus Type 1. Columns from left to right: (1) the identification of the sequence/strain, (2) the number of access to the NCBI base, (3) the patient from which it is coming the sequence, (4) the possible local of contamination of the patient.

Strain	Accession number	Patient ID	Infected area
D1/Hu/Saitama/NIID100/2014	LC011945	14-100J	Yoyogi Park, Tokyo
D1/Hu/Tokyo/NIID111/2014	LC011946	14-111J	Near Yoyogi Park, Tokyo
D1/Hu/Tokyo/NIID149/2014	LC011947	14-149J	Yotsuya-Shinjuku on the train, Tokyo
D1/Hu/Chiba/NIID153/2014	LC011948	14-153J	Chiba
D1/Hu/Shizuoka/NIID181/2014	LC011949	14-181J	Shizuoka?
D1/Hu/Hyogo/NIID188/2014	LC016760	14-188J	Hyogo? Malaysia?

then, the alphabet is $A = \{a, c, g, t\}$ with cardinal $|A| = 4$ and elements: adenine (a), cytosine (c), guanine (g) and thymine (t). All the sequences have around a size of 10 700 elements.

In Figure 1 we see a map of Japan with the regions from are coming the patients listed in Table 1. To calculate the classification of the sequences and establish the similarity between them, in the next section we first calculate the values of d_s for each pair of sequences, where s is a state of the state space $\mathcal{S}$. As usual, the elements of the alphabet A are organized in triples, then we can choose a memory $o = 3, 6, 9$, etc., therefore, the state space is composed of o concatenations of elements of A ($\mathcal{S} = A^o$). In this case, the size of the sequences is approximately 10 700, so the recommended memory is $o < \left\lfloor \log_{|A|}(10700) \right\rfloor - 1 = 7$, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x. Then we can use memory three or six, to simplify, we use the memory $o = 3$.

3.2 Similarity between the genomic sequences

Since we want to obtain global measurements between the sequences we calculate the values of $d_{\max}$ (Definition 2.1-ii.) between each pair of sequences. From this, we found that the three Tokyo sequences, LC011945, LC11946 and LC11947 have $d_{\max} = 0$. So, the three Tokyo sequences will be represented by LC011945. Then, we will work with four sequences, the sequence names an index number are shown in Table 2. Table 3 shows the value of $d_{\max}$ for each pair of sequences. That is, for each pair of sequences in the Table 2 we compute $d_{\max}$, the computation of $d_{\max}$ requires the computation of d_s for each s of the state space. And in that case the memory used is $o = 3$.

We see that the lowest value of $d_{\max}$ is caused by the sequences LC011945 and LC011948, with the second lowest value being the $d_{\max}$ between the sequences LC011945 and LC016760. Already the highest value of $d_{\max}$ occurs between the sequence LC011949 in relation to the sequences LC011945, LC011948 and LC016760 respectively. It is useful to represent the values of $d_{\max}$ through a dendrogram as seen on Figure 2. We build different dendrograms (average, median, single and complete) and they all point to the same organization between the sequences, see http://www.ime.unicamp.br/~jg/cadvj/. As we can see, in fact the dendrogram exposes the homogeneity between three of the four sequences: LC011945, LC011948 and LC016760, leaving exposed the disparity between the sequence LC011949 and the group of three sequences.

Observe that $d_{\max} < 1$ in all cases (Table 3), this implies that all values of $d_s < 1$ in all states $s \in \mathcal{S}$, i.e. the four sequences are considered as generated by the same stochastic law, but between them exist certain heterogeneity, detected by the magnitudes of $d_{\max}$. This fact allows us to carry out investigations that answer which of them is more or less representative in the set, which is the approach of the following subsection.

3.3 Classification of each sequence by means of V

We determine the classification attributed to each sequence, according to criterion V (see Definition 2.2). Table 4 shows the results.

The sequence that best represent de set of sequences (listed in Table 2) is coming from Tokyo LC011945. The most discrepant sequence (larger V) is LC011949, being the less representative sample and it indicates that LC011949 may have a different origin than the other sequences. This is, patient 14-181J was probably infected by a different strain from the other Japanese patients of Table 1. Comparing the classification of LC011949, which is 0.04200, we see clearly the impact of the $d_{\max}$ values coming from Table 3. Each time the sequence LC011949 is compared with another one in the list, the value of $d_{\max}$ increases by one decimal.

Figure 1. Map of Japan with the regions listed in Table 1.

Table 2. Genomic sequences and index number. The original set of six genomic sequences (Table 1) was reduced to four genomic sequences.

Index	Sequence name
1	LC011945 (representing LC011945, LC11946, and LC11947)
2	LC011948
3	LC011949
4	LC016760

Table 3. $d_{\max}$ values (see Definition 2.1-ii.). Columns 1 and 2 list the combinations of two sequences, from Table 2. Column 3 shows the value of $d_{\max}$ for the sequences to its left.

Sequence 1	Sequence 2	$d_{\max}$
LC011945	LC011948	0.00058
LC011945	LC011949	0.04204
LC011945	LC016760	0.00145
LC011948	LC011949	0.04204
LC011948	LC016760	0.00164
LC011949	LC016760	0.04204

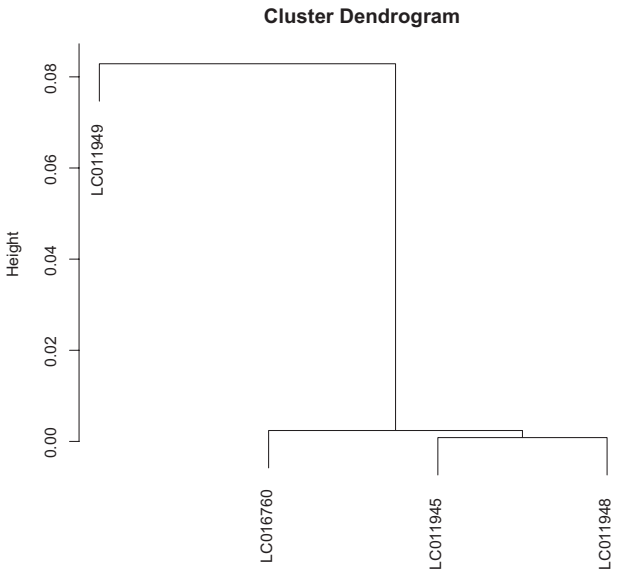


Figure 2. Dendrogram by *average* criterion build from the $d_{\max}$ values, reported in the Table 3.

Table 4. Value of V (Definition 2.2) for each sequence (see Table 2), ordered by increasing magnitude from top to bottom. In bold letter the most representative sequence (top) and the least representative sequence (bottom).

Sequence	Classification (V value)
LC011945	0.00145
LC011948	0.00164
LC016760	0.00164
LC011949	0.04200

To identify more strongly the meaning of this classification, we have compared all the sequences found in the base http://www.ncbi.nlm.nih.gov/ with the profile of being complete sequences of Dengue Type 1, year 2014, and coming from Asia. The list of accession numbers is given in the Table 5. For each sequence identified through its accession number we attach two letters to that number, in order to easily identify the country. In Figure 3 we show a dendrogram build by the *average* criterion with all the complete sequences.

Table 5. List of accession numbers (NCBI base) of complete sequences of Dengue virus Type 1, year 2014 – from Asia. The first column shows the country and the second column shows the sequences coming from the country, on the left.

Origin	Accession number of complete sequences
China (Ch)	KP686070.Ch, KP723473.Ch, KP723476.Ch, KR024705.Ch, KR024706.Ch, KR024707.Ch, KR024708.Ch, KR028435.Ch, KT187559.Ch, KT187560.Ch, KT187561.Ch, KT187562.Ch, KT187563.Ch, KT187564.Ch, KT827371.Ch, KT827372.Ch, KT827373.Ch, KT827374.Ch, KT827375.Ch, KT827376.Ch, KT827377.Ch, KT827378.Ch, KT827379.Ch, KU094071.Ch, KX225483.Ch, KX225484.Ch, KX225487.Ch, KX225488.Ch, KX225489.Ch, KX225490.Ch, KX225491.Ch, KX225492.Ch, KX225493.Ch, KX458013.Ch, KX458014.Ch, KX459386.Ch, KX459387.Ch, KX459388.Ch, KX459389.Ch, KX459390.Ch, KX459391.Ch, KX459392.Ch, KX620451.Ch, KX620452.Ch, KX620453.Ch, KX620454.Ch, KX620455.Ch, KX621249.Ch, KX621250.Ch, KX621252.Ch, MG560265.Ch, MG560266.Ch, MG560267.Ch, MG560268.Ch, MG560269.Ch
India (In)	KX618705.In, KX618706.In, KT831765.In
Japan (Ja)	LC011945.Ja, LC011948.Ja, LC011949.Ja, LC016760.Ja
Malaysia (Ma)	KU666939.Ma, KU666940.Ma, KU666941.Ma, KU666942.Ma, KX452050.Ma, KX452051.Ma, KX452052.Ma, KX452053.Ma, KX452054.Ma, KX452055.Ma, KX452056.Ma, KX452057.Ma, KX452058.Ma, KX452059.Ma, KX452060.Ma, KX452061.Ma, KX452062.Ma, KX452063.Ma, KX452064.Ma, KX452065.Ma, KX452067.Ma, KX452068.Ma
Singapore (Si)	KJ806959.Si, KJ806961.Si, KJ806963.Si, KX224261.Si, KX224263.Si, KY921903.Si, MF033213.Si, MF033214.Si, MF033215.Si, MF033216.Si, MF033217.Si, MF033218.Si, MF033219.Si, MF033220.Si, MF033221.Si, MF033222.Si, MF033223.Si, MF033224.Si, MF033225.Si, MF033226.Si, MF033227.Si
Sri Lanka (Sr)	KP398852.Sr
Thailand (Ta)	KU365900.Ta

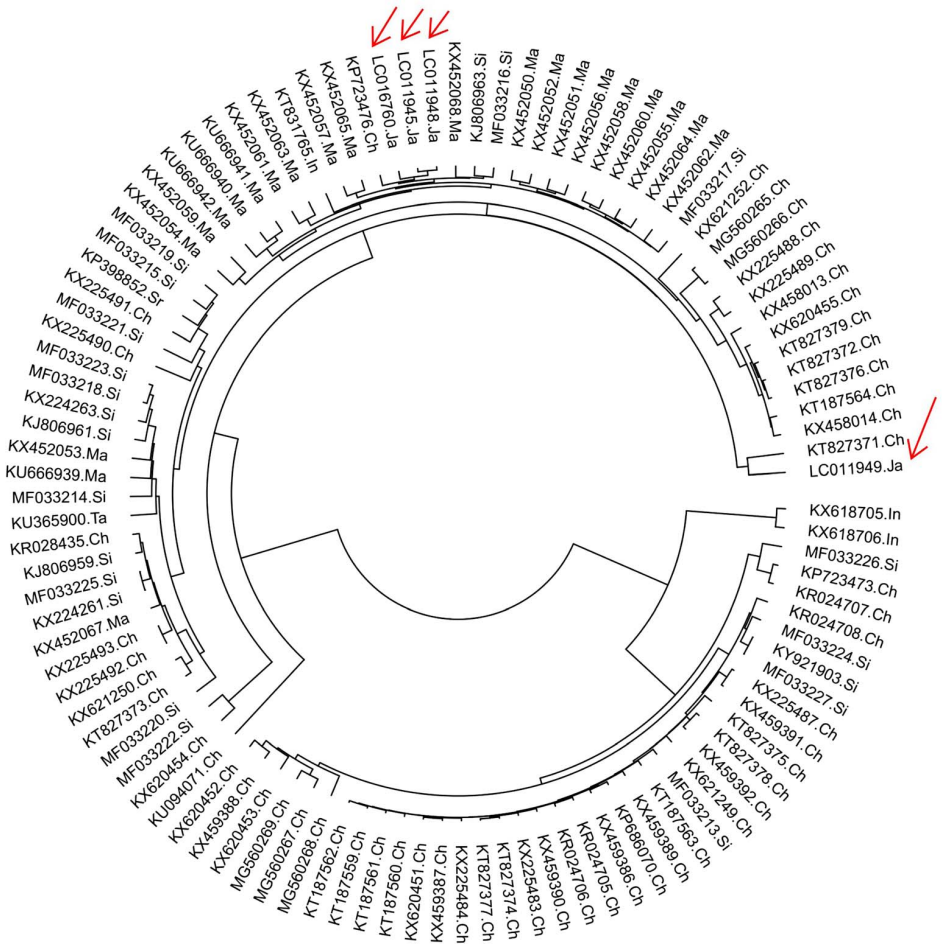


Figure 3. Dendrogram by *average* criterion for the sequences listed in Table 5 build from the $d_{\max}$ values (Definition 2.1-ii.). With arrows we indicate the four Japanese sequences from Table 2.

The circular dendrogram shows that the sequence of Japan (LC011949) with the highest V (among those in the list of Table 4) is in a cluster quite far away from the others in the list (LC011945, LC011948, LC016760). Some observations from Figure 3 can be done, for instance, the Japanese sequence LC011949 is next to the Chinese sequence KT827371, while the other Japanese sequences (Table 4) are closer to a variety of sequences from various countries including Japan, Malaysia and Singapore. Moreover, by the form of organization of the dendrogram, we verified that the sequence LC011949 is considerably more distant from the group {LC011945, LC011948, LC016760} in comparison with other foreign sequences, such as sequences coming from China, Malaysia and Singapore. As seen in Figure 2, the dendrogram of Figure 3 also shows the proximity of the Japanese sequences LC011945 and LC011948, which supports the argument of its representativeness in the group of Table 1. See also http://www.ime.unicamp.br/~jg/cadvj/, in order to corroborate the results with dendrograms build applying several criteria. The Japanese sequence LC011949 (Shizuoka patient who never visited Yoyogi Park) besides being the least representative (V and $d_{\max}$ higher) is also shown in Figure 3 closer to those of Chinese origin, which could implies a contamination of different origin.

4 Conclusion

In this paper we use two stochastic and statistically consistent notions to, (i) establish the proximity between genomic sequences (see [2]), (ii) classify the sequences in terms of their representativeness (see [1]). The classification rule gives low values to more representative sequences and it gives high values to less representative sequences. We classify genomic sequences of Dengue Virus Type I, originating in Japan and all corresponding to the outbreak occurred in Japan during 2014 (see Table 4). We identify the most representative sequences of the outbreak (those are from Tokyo), and we verify that these resemble other sequences (of 2014) coming from countries like Malaysia, Singapore, and China. The less representative sequence of the outbreak (from Shizuoka) is also a sequence that could resemble another one of Chinese origin (from 2014), but the latter being distant from the representative sequences of the outbreak. According to the classification that we have obtained and because of the evidence (see Figs. 2 and 3) we tend to agree with [3] in the sense of affirming that the outbreak in Japan during 2014 could involve more than one type of Dengue Virus Type I. By means of this type of approach it is possible to quantify the representativity of sequences, when compared with groups of sequences. This way of classifying is a genuinely stochastic tool, as explained in Section 2, that reports how close or distant are the stochastic laws of the sequences under consideration.

Future research could include the various serotypes of Dengue virus, in order to, (a) establish whether the notion $d_{\max}/d_s$ is capable of discriminating between the serotypes, (b) identify the spectrum of variation of the classifier (V) in each serotype, (c) establish the impact of the α constant (see Definition 2.1) in (a) and (b).

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. Fernández M, García Jesús E, Gholizadeh R, González-López VA (2019), Sample selection procedure in daily trading volume processes. Math Meth Appl Sci, 1–13. https://doi.org/10.1002/mma.5705.
2. 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.
3. Tajima S, Nakayama E, Kotaki A, Moi ML, Ikeda M, Yagasaki K, Saito Y, Shibasaki K, Saijo M, Takasaki T (2017), Whole genome sequencing–based molecular epidemiologic analysis of autochthonous dengue virus type 1 strains circulating in Japan in 2014. Jpn J infect Dis 70, 1, 45–49.
4. 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.
5. 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.
6. 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.

7. García Jesús E, González-López VA (2016), Markov partition models for Epstein Barr virus, in: JR Bozeman Jr, T Oliveira, CH Skiadas (Eds.), Stochastic and Data Analysis Methods and Applications in Statistics and Demography, International Society for the Advancement of Science and Technology (ISAST), Athens.
8. García Jesús E, Gholizadeh R, González-López VA (2018), Stochastic distance between Burkitt lymphoma/leukemia strains, in: C Skiadas, C Skiadas (Eds.), Demography and Health Issues. The Springer Series on Demographic Methods and Population Analysis, Vol. 46, Springer, Cham.
9. Baer R, Bankier AT, Biggin MD, Deininger PL, Farrell PJ, Gibson TJ, Hatfull G, Hudson GS, Satchwell SC, Seguin C, Tuffnell PS, Barrell BG (1984), DNA sequence and expression of the B95–8 Epstein-Barr virus genome. Nature 310, 5974, 207.
10. Schwarz G (1978), Estimating the dimension of a model. Ann Stat 6, 2, 461–464.

Cite this article as: Cordeiro MTA, García JE, González-López VA & Londoño SLM 2019. Classification of autochthonous dengue virus type 1 strains circulating in Japan in 2014. 4open, **2**, 20.

D.2 Stochastic profile of Epstein-Barr virus in nasopharyngeal carcinoma settings

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 $a_m, a_{m+1},\ldots, a_n$ 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 $\left\{x_1^{J\,n_j}\right\}_{j=1}^p$ be samples of the processes $\{(X_t^j)_{t\in\mathbb{Z}}\}_{j=1}^p$, with sample sizes $\{n_j\}_{j=1}^p$. Then, $\left\{x^{j n_j}_1\right\}_{j=1}^p$ 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 $\{x_1^{l^{n_l}}\}_{l=1}^{p}$ 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.