4open 2019, **2**, 20
© M.T.A. Cordeiro et al., Published by EDP Sciences, 2019
https://doi.org/10.1051/fopen/2019018

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

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

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*

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

where, given a sequence $\{z_j\}_{j=1}^{l}$*,* $median\,\{z_j, 1 \leq j \leq l\} = z_{(k+1)}$ *if* $l = 2k + 1$ *and* $median\,\{z_j, 1 \leq j \leq l\} = \frac{z_{(k)}+z_{(k+1)}}{2}$ *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)

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

(ii)

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

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.

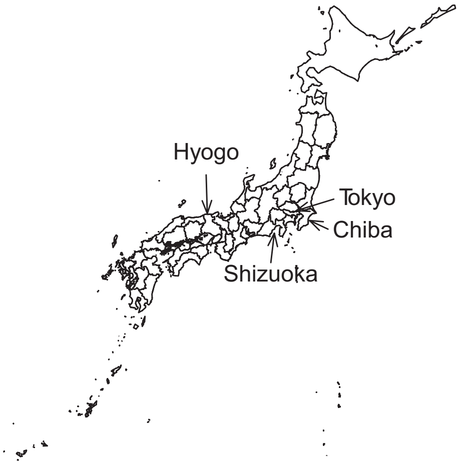

Figure 1. Map of Japan with the regions listed in Table 1.

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

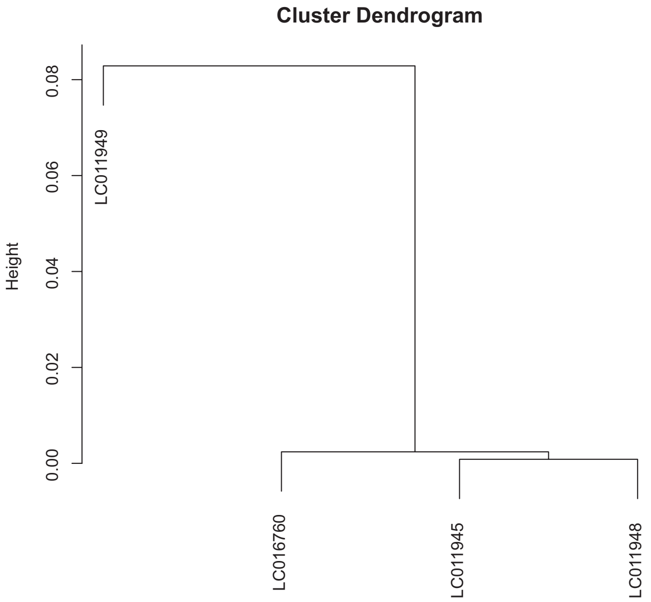

Figure 2. Dendrogram by *average* criterion build from the $d_{\max}$ values, reported in the Table 3.

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

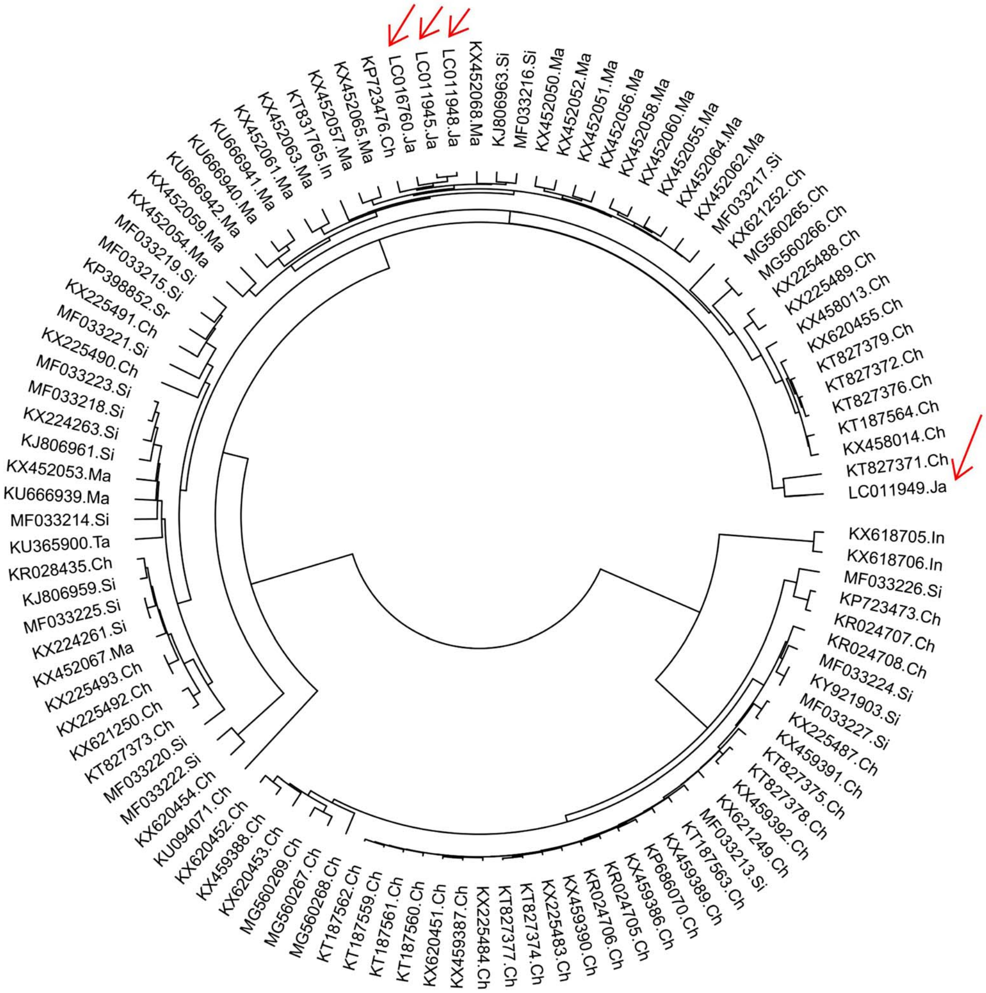

Figure 3. Dendrogram by *average* criterion for the sequences listed in Table 5 build from the $d_{\max}$ values (Definition 2.1-ii.). With arrows we indicate the four Japanese sequences from Table 2.

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.