Comparação Estocástica Entre a Estrutura Genética Original do SARS-CoV-2 e a Variante SARS-CoV-2 - P.1

Jesús E. García, V. A. González-López, e G. H. Tasca

5.1 Introdução

As variantes do vírus SARS-CoV-2 começaram a se espalhar globalmente no final de 2020 e início de 2021, desafiando as vacinas originalmente criadas para o vírus em suas versões de 2019/2020. Isso levou à necessidade de modificações nas fórmulas das vacinas ou na administração, como o ajuste no número de doses para alcançar os resultados esperados. Em alguns locais, variantes recentes passaram a dominar as infecções. Por exemplo, estima-se que no estado de São Paulo, Brasil, a variante P.1 (identificada originalmente no Amazonas) seja responsável por aproximadamente 90% das contaminações. Em abril de 2021, o Brasil enfrentou o mês mais letal de toda a pandemia de SARS-CoV-2, e especialistas apontam que isso pode ser atribuído à predominância da variante P.1. Devido à falta de testagem em massa, essa estimativa é aproximada.

Variantes de um vírus são fenômenos esperados, já que mutações na estrutura genética de um vírus ocorrem naturalmente. Essas mutações podem resultar em versões menos ou mais infecciosas, mas não é determinístico que as mutações levem a variantes mais contagiosas. O site `https://cov-lineages.org/llama.html` exibe um mapa com as variantes do vírus SARS-CoV-2. A preocupação está relacionada às variantes que conseguem evadir eficientemente o sistema imunológico humano. Esse é o caso de três variantes recentes do SARS-CoV-2: B.1.1.7 (Reino Unido), B.1.351 (África do Sul) e P.1 (Brasil), conforme Hoffmann et al. (2021).

Determinar se uma mutação resulta em uma variante é um desafio, pois isso depende da localização da mutação. Por exemplo, sinais de variações do vírus já haviam sido identificados durante o primeiro semestre de 2020 (veja García et al., 2020). Entretanto, algumas variantes possuem elevado poder de contaminação, como já foi determinado no caso da variante P.1.

Em García et al. (2020), foi analisado o primeiro registro genético do vírus SARS-CoV-2, uma sequência completa do genoma coletada em dezembro de 2019. Essa sequência foi extraída de um paciente vindo do mercado de frutos do mar de Wuhan, associado à origem do surto. O número de acesso da sequência é MN908947 (versão MN908947.3), disponível em `https://www.ncbi.nlm.nih.gov/nuccore/MN908947`, veja também Wu et al. (2020). García et al. (2020) exploraram o formato FASTA de MN908947, composto por 29903 bases: `a`, `c`, `g`, `t`. Cordeiro et al. (2020) sugerem que essa sequência, considerada como uma amostra de um processo estocástico no alfabeto $\mathcal{A} = \{a, c, g, t\}$, pode ser bem representada por um Modelo de Partição de Markov (Partition Markov Model - PMM) com uma estrutura peculiar, introduzida em García et al. (2020).

Os Modelos de Partição de Markov foram originalmente introduzidos em García e González-López (2011) (veja também García e González-López, 2017) e mostraram ser ferramentas poderosas para analisar e representar o comportamento estocástico de sequências genéticas (veja, por exemplo, Cordeiro et al., 2020).

Os objetivos deste trabalho são: (1) investigar o comportamento de uma coleção de sequências da variante SARS-CoV-2 P.1 sob a perspectiva dos Modelos de Partição de Markov; e (2) comparar a sequência MN908947 com o conjunto de sequências da variante SARS-CoV-2 P.1 utilizando ferramentas estocásticas associadas aos Modelos de Partição de Markov, a fim de identificar as alterações ocorridas no vírus em termos de sua representação estocástica.

Este artigo está organizado da seguinte forma: a Seção 5.2 apresenta as ferramentas estocásticas utilizadas para a comparação das sequências genéticas. A Seção 5.3 introduz a base de dados e apresenta os resultados das comparações e dos modelos obtidos. A Seção 5.4 contém nossas conclusões.

5.1 Introdução

As variantes do vírus SARS-CoV-2 começaram a se espalhar globalmente no final de 2020 e início de 2021, desafiando as vacinas originalmente criadas para o vírus em suas versões de 2019/2020. Isso levou à necessidade de modificações nas fórmulas das vacinas ou na administração, como o ajuste no número de doses para alcançar os resultados esperados. Em alguns locais, variantes recentes passaram a dominar as infecções. Por exemplo, estima-se que no estado de São Paulo, Brasil, a variante P.1 (identificada originalmente no Amazonas) seja responsável por aproximadamente 90% das contaminações. Em abril de 2021, o Brasil enfrentou o mês mais letal de toda a pandemia

de SARS-CoV-2, e especialistas apontam que isso pode ser atribuído à predominância da variante P.1. Devido à falta de testagem em massa, essa estimativa é aproximada.

Variantes de um vírus são fenômenos esperados, já que mutações na estrutura genética de um vírus ocorrem naturalmente. Essas mutações podem resultar em versões menos ou mais infecciosas, mas não é determinístico que as mutações levem a variantes mais contagiosas. O site `https://cov-lineages.org/llama.html` exibe um mapa com as variantes do vírus SARS-CoV-2. A preocupação está relacionada às variantes que conseguem evadir eficientemente o sistema imunológico humano. Esse é o caso de três variantes recentes do SARS-CoV-2: B.1.1.7 (Reino Unido), B.1.351 (África do Sul) e P.1 (Brasil), conforme Hoffmann et al. (2021).

Determinar se uma mutação resulta em uma variante é um desafio, pois isso depende da localização da mutação. Por exemplo, sinais de variações do vírus já haviam sido identificados durante o primeiro semestre de 2020 (veja García et al., 2020). Entretanto, algumas variantes possuem elevado poder de contaminação, como já foi determinado no caso da variante P.1.

Em García et al. (2020), foi analisado o primeiro registro genético do vírus SARS-CoV-2, uma sequência completa do genoma coletada em dezembro de 2019. Essa sequência foi extraída de um paciente vindo do mercado de frutos do mar de Wuhan, associado à origem do surto. O número de acesso da sequência é MN908947 (versão MN908947.3), disponível em `https://www.ncbi.nlm.nih.gov/nuccore/MN908947`, veja também Wu et al. (2020). García et al. (2020) exploraram o formato FASTA de MN908947, composto por 29903 bases: `a`, `c`, `g`, `t`. Cordeiro et al. (2020) sugerem que essa sequência, considerada como uma amostra de um processo estocástico no alfabeto $\mathcal{A} = \{a, c, g, t\}$, pode ser bem representada por um Modelo de Partição de Markov (Partition Markov Model - PMM) com uma estrutura peculiar, introduzida em García et al. (2020).

Os Modelos de Partição de Markov foram originalmente introduzidos em García e González-López (2011) (veja também García e González-López, 2017) e mostraram ser ferramentas poderosas para analisar e representar o comportamento estocástico de sequências genéticas (veja, por exemplo, Cordeiro et al., 2020).

Os objetivos deste trabalho são: (1) investigar o comportamento de uma coleção de sequências da variante SARS-CoV-2 P.1 sob a perspectiva dos Modelos de Partição de Markov; e (2) comparar a sequência MN908947 com o conjunto de sequências da variante SARS-CoV-2 P.1 utilizando ferramentas estocásticas associadas aos Modelos de Partição de Markov, a fim de identificar as alterações ocorridas no vírus em termos de sua representação estocástica.

Este artigo está organizado da seguinte forma: a Seção 5.2 apresenta as ferramentas estocásticas utilizadas para a comparação das sequências genéticas. A Seção 5.3 introduz a base de dados e apresenta os resultados das comparações e dos modelos obtidos. A Seção 5.4 contém nossas conclusões.

1 Modelos e Critérios

Nesta seção, introduzimos as noções de uma cadeia de Markov com partição, denominada como Modelo de Partição de Markov (PMM). Consideraremos cada uma das sequências genômicas como uma amostra de um PMM. Em seguida, introduzimos o conceito que nos permite estabelecer a similaridade/dissimilaridade entre as sequências do SARS-CoV-2. As seguintes noções são apresentadas para identificar o melhor modelo, sequência por sequência, para o conjunto de sequências completas da variante SARS-CoV-2 - P.1 dentro de uma família de modelos onde a sequência original MN908947 foi melhor representada.

Também introduzimos uma métrica que permite quantificar a similaridade/dissimilaridade entre as sequências da P.1 e entre as sequências da P.1 e a sequência original MN908947. Utilizando os parâmetros indicados (identificados por um procedimento de seleção de modelos), medimos a similaridade/dissimilaridade entre as sequências e identificamos os estados que produzem as discrepâncias.

Seja (X_t) uma cadeia de Markov de tempo discreto de ordem o (o finito) em um alfabeto finito A e G um valor tal que $G > o$. Denote por $S = A \times A^o$ o espaço de estados do processo (X_t). Considere a notação a_k^n como a concatenação dos elementos $a_k a_{k+1} \dots a_n$, onde $a_i \in A, \forall i : k \leq i \leq n$. Os parâmetros que definem o comportamento do processo são as probabilidades de transição, introduzidas aqui:

$$P(a \mid (z,s)) = Prob(X_t = a \mid X_{t-G} = z, X_{t-1}^{t-o} = s), \tag{1}$$

para cada $a \in A$ e $(z,s) \in S$. Assim, precisamos identificar o conjunto de probabilidades que representam o comportamento do processo:

$$\{P(a \mid (z,s)), a \in A, (z,s) \in S\}.$$

Definição 2.1

Uma cadeia de Markov G-Markov (X_t) é uma cadeia de Markov de tempo discreto em um alfabeto finito A, com espaço de estados $S = A \times A^o$, onde

$o < \infty$, probabilidades de transição seguindo a Equação (5.1), para um G finito, tal que $G > o$.

Para estimar as probabilidades de transição, conforme a Definição 2.1, considere x_1^n uma amostra do processo (X_t), $(z, s) \in S$, $a \in A$ e $n > G$. Denote por $N((z, s))$ o número de ocorrências do estado (z, s) em x_1^n, ou seja, $N((z, s)) = |\{t : G < t \leq n, x_{t-G} = z, x_{t-1}^{t-o} = s\}|$ e as ocorrências de $(z, s) \in A \times A^o$ seguidas por $a \in A$ são $N((z, s), a) = |\{t : G < t \leq n, x_{t-G} = z, x_{t-1}^{t-o} = s, x_t = a\}|$. Assim, para cada $a \in A$ e $(z, s) \in S$, $\frac{N((z,s),a)}{N((z,s))}$ é o estimador de $P(a \mid (z, s))$ dado pela Equação (5.1),

$$\hat{P}(a \mid (z, s)) = \frac{N((z, s), a)}{N((z, s))}, \quad a \in A, (z, s) \in S. \tag{2}$$

Para obter uma estimativa eficiente, introduzimos um modelo que reduz o número de probabilidades a serem estimadas. A ideia é usar diferentes estados para estimar a mesma probabilidade.

Definição 2.2

Seja (X_t) uma cadeia de Markov seguindo a Definição 2.1, de ordem o em um alfabeto finito A, com parâmetro $G > o$ e espaço de estados $S = A \times A^o$:

i. $v, r \in S$ são equivalentes se $P(a \mid v) = P(a \mid r), \forall a \in A$.

ii. (X_t) é uma cadeia G-Markov com partição $I = \{I_1, I_2, \ldots, I_{|I|}\}$ se esta partição for definida pela relação introduzida no item i.

Seja $I = \{I_1, I_2, \ldots, I_{|I|}\}$ uma partição de S, definimos as probabilidades em termos de partes (de I). Seja $P(I, a) = \sum_{r \in I} P(r, a)$ e $P(I) = \sum_{r \in I} P(r)$. Se $P(I) > 0$, podemos definir:

$$P(a \mid I) = \frac{P(I, a)}{P(I)}. \tag{3}$$

Assim, $\forall a \in A$, $P(a \mid I) = P(a \mid r), \forall r \in I$, significando que usamos todos os estados $r \in I$ para estimar o mesmo parâmetro. Essa forma de representar um processo estocástico (X_t) é chamada de Modelo de Partição de Markov (PMM), veja García e González-López (2017) e García et al. (2020).

A estimativa do modelo proposto na Definição 2.2 é realizada maximizando o Critério de Informação Bayesiano (BIC), veja Schwarz (1978). Defina o BIC como:

$$BIC(x_1^n, I) = \ln \left(\prod_{a \in A, I \in I} \left(\frac{N(I, a)}{N(I)} \right)^{N(I,a)} \right) - (|A| - 1)|I| \ln(n)\alpha. \tag{4}$$

2 SARS-CoV 2 e Variante P.1

Nesta seção, descrevemos, na Seção 5.3.1, as sequências de SARS-CoV 2 - variante P.1 que são investigadas sob as noções introduzidas na Seção 5.2. Na Seção 5.3.2, apresentamos os modelos selecionados para cada sequência introduzida na Seção 5.3.1. Para identificar evidências de mudanças na sequência original de SARS-CoV 2 e no conjunto de sequências da variante SARS-CoV 2 - P.1, comparamos essas sequências utilizando as ferramentas apresentadas na Seção 5.2.

2.1 Conjuntos de Dados do SARS-CoV 2

O banco de dados consiste em uma coleção de sequências genéticas no formato FASTA. Por essa razão, o alfabeto considerado é o genômico, ou seja, $\mathcal{A} = \{a, c, g, t\}$. As sequências completas do genoma da variante SARS-CoV 2 - P.1 utilizadas neste artigo podem ser encontradas na fonte GISAID (`https://gisaid.org`), com as sequências listadas na Tabela 1.

A Tabela 1 registra o ID de acesso (*Accession ID*) de cada sequência, a data de coleta (janeiro de 2021) e o tamanho da amostra (pelo menos 29.593). O laboratório de origem das sequências EPI_ISL_1034306 e EPI_ISL_106828x para $x = 1, 2, 3, 4$ é o *Laboratório de Ecologia de Doenças Transmissíveis na Amazônia, Instituto Leonidas e Maria Deane - Fiocruz, Amazônia.* O laboratório de origem da sequência EPI_ISL_906071 é o *LACEN - Laboratório Central de Saúde Pública Dr. Costa Alvarenga, Piauí.* O laboratório de origem das sequências EPI_ISL_90608x para $x = 0, 1$ é o *Hospital Beneficência Portuguesa, São Paulo.* O laboratório de origem das sequências EPI_ISL_94062x para $x = 6, 7$ é o *Hospital Central São Caetano do Sul, São Paulo.*

Na subseção a seguir, apresentamos os resultados do ajuste do modelo da Definição 2.1 em cada sequência listada na Tabela 1, bem como um estudo comparativo que busca identificar a similaridade e divergência entre as sequências listadas na Tabela 1 com a sequência MN908947 (Wuhan).

3 SARS-CoV 2 e Variante P.1

Nesta seção, descrevemos, na Seção 5.3.1, as sequências do SARS-CoV 2 - variante P.1 que são investigadas sob as noções introduzidas na Seção 5.2. Na Seção 5.3.2, apresentamos os modelos selecionados para cada sequência introduzida na Seção 5.3.1. Para identificar evidências de mudanças em relação à sequência original do SARS-CoV 2 e ao conjunto de sequências

da variante SARS-CoV 2 - P.1, comparamos essas sequências utilizando as ferramentas introduzidas na Seção 5.2.

3.1 Conjuntos de Dados do SARS-CoV 2

A base de dados consiste em uma coleção de sequências genéticas no formato FASTA. Por essa razão, o alfabeto considerado é o genômico, ou seja, $A = \{a, c, g, t\}$. As sequências completas do genoma do SARS-CoV 2 - variante P.1 utilizadas neste trabalho podem ser encontradas na fonte GISAID (`https://gisaid.org`), e as sequências estão listadas na Tabela 2.

A Tabela 2 registra o ID de acesso de cada sequência, a data de coleta (janeiro de 2021) e os tamanhos das amostras (pelo menos 29.593). O laboratório de origem das sequências EPI_ISL_1034306, EPI_ISL_106828x para $x = 1, 2, 3, 4$ é o Laboratório de Ecologia de Doenças Transmissíveis na Amazônia, Instituto Leonidas e Maria Deane - Fiocruz, Amazônia. O laboratório de origem da sequência EPI_ISL_906071 é o LACEN - Laboratório Central de Saúde Pública Dr. Costa Alvarenga, Piauí. O laboratório de origem das sequências EPI_ISL_90608x para $x = 0, 1$ é o Hospital Beneficência Portuguesa, São Paulo. E o laboratório de origem das sequências EPI_ISL_94062x para $x = 6, 7$ é o Hospital Central São Caetano do Sul, São Paulo.

Na subseção a seguir, mostramos os resultados do ajuste do modelo da Definição 2.1 em cada sequência listada na Tabela 2 e um estudo comparativo que busca identificar similaridades e divergências entre as sequências listadas na Tabela 2 com a sequência MN908947 (Wuhan).

3.2 Resultados da Comparação de Sequências

A Tabela 3 apresenta a continuação da Tabela **??**. Da esquerda para a direita, cada tabela informa (1) a sequência com a qual a sequência no topo foi comparada, (2) o valor de d_{max} e (3) o estado onde d_{max} ocorre.

As comparações entre as sequências da variante P.1 e a sequência MN908947 mostram um aumento no valor de d_r, de aproximadamente 0.0134 para 0.1, como registrado na Tabela **??**. Esse aumento reflete a separação da sequência MN908947 em relação aos clusters da variante P.1. As transições entre os estados na Tabela **??** indicam que a probabilidade de transição do estado (a, aaa) para a diminui em cerca de 20%, enquanto para t aumenta em relação à sequência original de Wuhan.

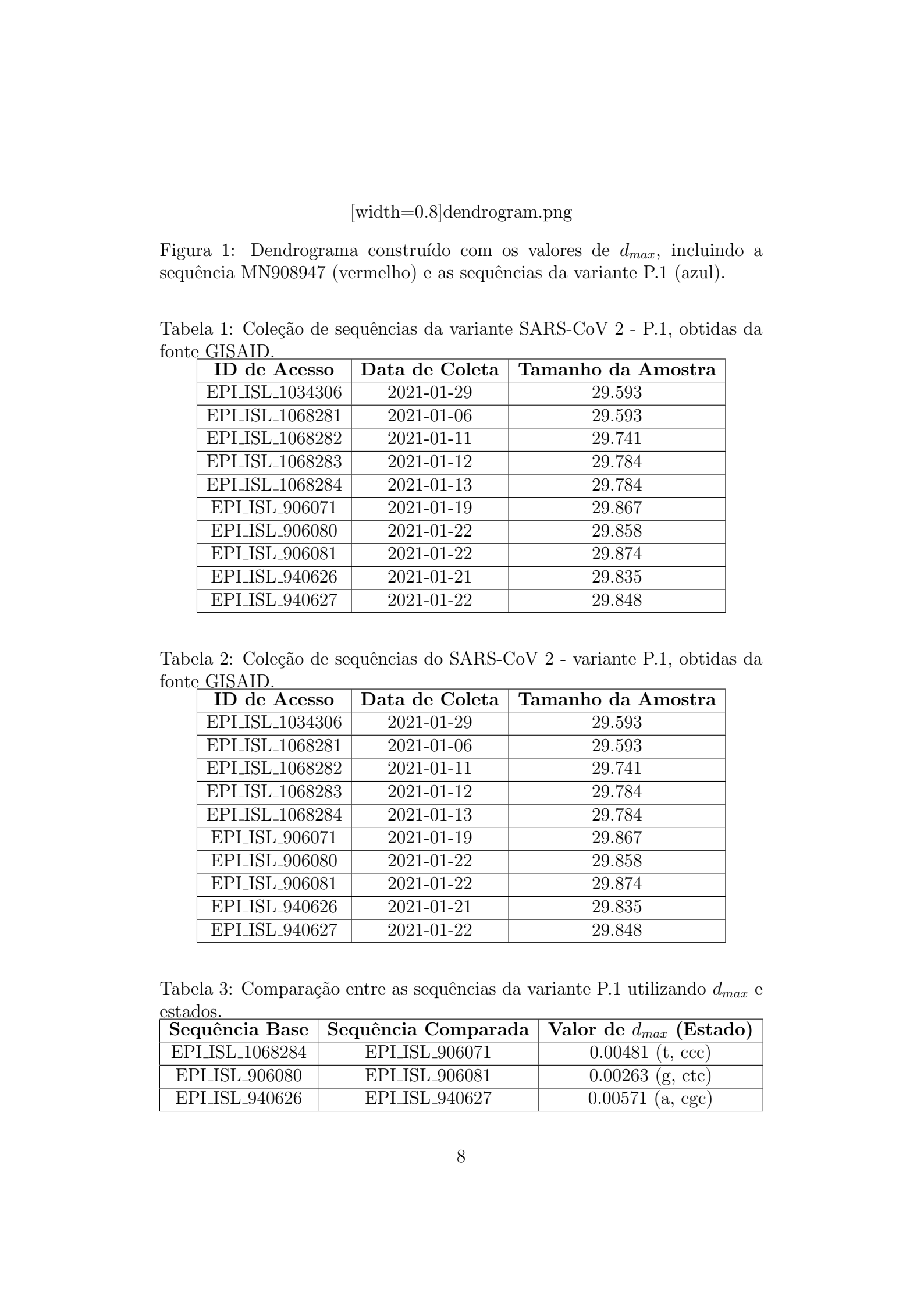

Figura 1: Dendrograma construído com os valores de d_{max}, incluindo a sequência MN908947 (vermelho) e as sequências da variante P.1 (azul).

Tabela 1: Coleção de sequências da variante SARS-CoV 2 - P.1, obtidas da fonte GISAID.

ID de Acesso	**Data de Coleta**	**Tamanho da Amostra**
EPI_ISL_1034306	2021-01-29	29.593
EPI_ISL_1068281	2021-01-06	29.593
EPI_ISL_1068282	2021-01-11	29.741
EPI_ISL_1068283	2021-01-12	29.784
EPI_ISL_1068284	2021-01-13	29.784
EPI_ISL_906071	2021-01-19	29.867
EPI_ISL_906080	2021-01-22	29.858
EPI_ISL_906081	2021-01-22	29.874
EPI_ISL_940626	2021-01-21	29.835
EPI_ISL_940627	2021-01-22	29.848

Tabela 2: Coleção de sequências do SARS-CoV 2 - variante P.1, obtidas da fonte GISAID.

ID de Acesso	**Data de Coleta**	**Tamanho da Amostra**
EPI_ISL_1034306	2021-01-29	29.593
EPI_ISL_1068281	2021-01-06	29.593
EPI_ISL_1068282	2021-01-11	29.741
EPI_ISL_1068283	2021-01-12	29.784
EPI_ISL_1068284	2021-01-13	29.784
EPI_ISL_906071	2021-01-19	29.867
EPI_ISL_906080	2021-01-22	29.858
EPI_ISL_906081	2021-01-22	29.874
EPI_ISL_940626	2021-01-21	29.835
EPI_ISL_940627	2021-01-22	29.848

Tabela 3: Comparação entre as sequências da variante P.1 utilizando d_{max} e estados.

Sequência Base	**Sequência Comparada**	**Valor de d_{max} (Estado)**
EPI_ISL_1068284	EPI_ISL_906071	0.00481 (t, ccc)
EPI_ISL_906080	EPI_ISL_906081	0.00263 (g, ctc)
EPI_ISL_940626	EPI_ISL_940627	0.00571 (a, cgc)