Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

Caracterização e visualização da diversidade genética do vírus Influenza ao longo do tempo

Rafaela Gomes de Jesus

18 de dezembro de 2018

Rafaela Gomes de Jesus

Caracterização e visualização da diversidade genética do vírus Influenza ao longo do tempo

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientadora: Profa. Dr. Gabriela Bettela Cybis

Porto Alegre
Novembro de 2018

Rafaela Gomes de Jesus

Caracterização e visualização da diversidade genética do vírus Influenza ao longo do tempo

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pela Orientadora e pela Banca Examinadora.

Orientadora:_____
Profa. Dr. Gabriela Bettela Cybis, UFRGS
Doutora pela Universidade da Califórnia, Los Angeles, CA

Banca Examinadora:

Prof. Dr. Álvaro Vigo, UFRGS
Doutor pela Universidade Federal do Rio Grande do Sul – Porto Alegre, RS

Prof. Dr. Cleber Bisognin, UFRGS
Doutor pela Universidade Federal do Rio Grande do Sul – Porto Alegre, RS

Porto Alegre
Novembro de 2018

Agradecimentos

Aos meus pais, Adriana e Jeferson, por todo amor e apoio que sempre me deram e por serem os responsáveis por tudo isso ser possível. À minha irmã, Júlia, que é minha eterna companheira e me enche de orgulho. Aos meus tios, tias, primos, primas e minhas duas avós que sempre torceram por mim e se alegram com todas as minhas conquistas. Às minhas amigas, irmãs de coração, Marina, Roberta e Shakira que, além de dividirem o apartamento, compartilharam a vida comigo. Aos professores do Departamento de Estatística por todos os ensinamentos, em especial à professora Gabriela, pela orientação neste trabalho e por todo o conhecimento transmitido. À UFRGS, por me proporcionar experiências e amigos incríveis ao longo desses quatro anos de graduação.

Resumo

A gripe é uma doença que se propaga globalmente todos os anos, infectando de 10 a 20% da população mundial, causando mortes e grandes perdas econômicas. Isso acontece porque o vírus Influenza possui altas taxas de mutação em seu material genético, o que é fator essencial para o seu sucesso epidemiológico. Por conta disso, a Organização Mundial da Saúde reúne esforços, todos os anos, para criar vacinas que ajudem a reduzir os casos da doença. Para direcionar, mais precisamente, as medidas preventivas que devem ser tomadas para que a doença seja evitada, é de grande importância monitorar as alterações genéticas dos vírus. Um forma de realizar o monitoramento é através das medidas de diversidade genética, que mensuram a variabilidade de uma população.

Este trabalho tem como objetivo investigar e caracterizar o comportamento ao longo do tempo de dois vírus Influenza tipo A, o H1N1 e o H3N2, e duas linhagens do vírus Influenza tipo B, o Victoria e o Yamagata. O primeiro passo para a realização deste estudo foi montar o banco de dados com as sequências genéticas dos vírus e fazer o seu pré-processamento. Posteriormente, foi necessário uma pesquisa sobre medidas de diversidade genética para escolher a melhor forma de analisar os dados.

A análise da evolução da diversidade genética foi realizada separadamente para cada tipo de vírus estudado neste trabalho. Para melhor compreensão do comportamento dos vírus ao longo do tempo, os resultados foram apresentados, em grande parte, de modo gráfico. Para o presente estudo, além da medidas de diversidade genética, fez-se uso de conhecimentos e técnicas estatísticas como a análise de agrupamentos e o escalonamento multidimensional e de programação em R.

Como resultado, obteve-se a representação gráfica das medidas de diversidade para cada tipo de vírus em série temporal por ano e em série temporal no esquema de janela deslizante, com três tamanhos de janela k=1 mês, 3 meses e 6 meses. Obteve-se também a representação gráfica da distância genética entre os vírus, através da técnica de escalonamento multidimensional. A partir das análises dos resultados obtidos, as principais conclusões deste estudo são que a diversidade genética dos vírus é melhor representada pela série temporal com janela deslizante k=*3*, tendo seus resultados condizentes com os gráficos do escalonamento multidimensional e que o nível de similaridade entre os vírus depende de sua distância no tempo, isso porque quanto maior o período entre vírus, mais sujeitos eles estão às mutações genéticas.

Palavras-Chave: Diversidade genética, Escalonamento Multidimensional, Diversidade genética, Influenza, Epidemiologia Genética.

Abstract

Influenza is a disease that spreads globally every year, infecting to 10 from 20% of the world's population, causing deaths and major economic losses. This happens because the Influenza virus has high rates of mutation in its genetic material, which is an essential factor for its epidemiological success. Therefore, the World Health Organization works every year to create vaccines to help reducing cases of this disease. To address the preventive measures that must be taken to prevent the disease, it is importante to monitor the genetic changes of the virus. One way to monitor is through genetic diversity measures, which tracks the variability of a population.

This work aims to investigate and to characterize the behavior over time of two type Influenza virus A, H1N1 and H3N2, and two lineages of influenza virus of type B, Victoria and Yamagata. The first step in this study was to assemble the database with the genetic sequences of the viruses and preprocess them. Subsequently, it was necessary to research different measures of genetic diversity to choose the best one to analyze the data.

Evolution analysis of genetic diversity was performed separately for each type of virus studied in this work. For a better understanding of virus behavior over time, the results were mostly presented through graphs. For the present study, in addition to genetic diversity measures, techniques such as cluster analysis, multidimensional scaling and R programming were used.

As a result, we obtained the graphical representation of the diversity measures for each type of virus in time series per year and in time series in sliding window scheme, with three window sizes $k = 1$ month, 3 months and 6 months. It was also obtained the graphical representation of the genetic distance between the viruses, through multidimensional scaling technique. From the analysis of the results, the conclusions of this study are that the genetic diversity of the virus is best represented by the time series in sliding window scheme with $k = 3$, and its results are consistent with the multidimensional scaling graphs and the level of similarity between viruses depends on their distance in time, because the longer the period between viruses, the more exposed they are to the genetic mutations.

Keywords: Genetic Diversity, Multidimensional Scaling, Biosequence Analysis, Influenza, Genetic Epidemiology.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

A gripe é um problema de saúde pública causado pelo vírus Influenza e que se propaga globalmente em ciclos sazonais de epidemias. A doença infecta de 10% a 20% da população mundial todos os anos, gerando aproximadamente 500.000 mortes e grandes perdas econômicas (Stöhr, 2002; Neher e Bedford, 2015). Os danos são ainda maiores quando ocorrem pandemias, elevando o número de óbitos a milhões. O sucesso epidemiológico do vírus que causa a doença se deve, principalmente, às altas taxas de mutação em seu material genético. Frequentemente, essas mutações resultam em novas variantes do vírus, que por sua vez infectam a população que ainda não tem imunidade para combatê-lo em sua versão modificada.

Na tentativa de proteger a população contra a gripe e suas consequências, a Organização Mundial da Saúde reúne esforços todos os anos para criar vacinas. Estas estimulam o sistema imunológico a se defender dos novos vírus circulantes e, dessa forma, ajudam a prevenir epidemias da doença. Contudo, o processo de produção da vacina requer tempo e, dada às altas taxas de mutação, a linhagem do vírus que circula no momento da sua fabricação não é a mesma da que circula no período de vacinação. Cabe aos especialistas fazer a previsão, baseando-se em dados monitorados e estatísticas, de qual será a linhagem predominante quando a vacina estiver disponível para uso.

A decisão do design da futura vacina é extremamente importante, pois caso não esteja de acordo com o vírus que circula no momento da imunização, pode gerar mortes e grandes prejuízos econômicos. Por isso, é de grande importância monitorar as alterações genéticas do vírus para conhecer sua dinâmica e direcionar, mais precisamente, as medidas preventivas que devem ser tomadas para que a doença seja evitada. Uma das formas de realizar o monitoramento é utilizar medidas de diversidade genética e ferramentas estatísticas que mensurem a dimensão da variação entre os vírus Influenza. Essas medidas podem ser sinalizadoras de padrões de mudanças dos vírus, podendo agregar informações para a previsão das futuras linhagens dominantes.

Neste trabalho serão estudados duas variantes do vírus influenza A: H1N1, H3N2; e duas variantes do Influenza B: Victoria e Yamagata. O objetivo do estudo é investigar e descrever, através de medidas de diversidade genética e técnicas estatísticas, aspectos moleculares da evolução dos vírus escolhidos para a análise. Para melhor e mais efetiva compreensão do comportamento dos vírus ao longo do tempo, os resultados analíticos deste trabalho são acompanhados de resultados gráficos.

O presente estudo envolve desde a montagem dos bancos de dados até a caracterização visual da evolução dos vírus H1N1, H3N2, Victoria e Yamagata. Para

realização deste trabalho, foi necessário unir conhecimentos de biologia, programação e estatística. Dentro de um escopo maior, este trabalho assume o papel de gerar informações pertinentes sobre os vírus Influenza que possam contribuir para um futuro estudo sobre a previsão de linhagens dominantes.

2 Referencial teórico

Neste capítulo serão abordados os conceitos centrais para o presente estudo, com uma breve revisão da literatura.

2.1 O vírus Influenza

O vírus Influenza é uma partícula esférica cuja superfície é coberta por proteínas com funções essenciais ao vírus: a hemaglutinina, responsável pela entrada do vírus nas células onde este se irá multiplicar; e a neuraminidase, que permite a libertação dos novos vírus que irão à conquista de novas células. O vírus da gripe apresenta um genoma constituído por segmentos de ácido ribonucleico, mais conhecido como RNA, que comanda o seu funcionamento (Taubenberger e Morens, 2008). O RNA de cada vírus consiste em oito cromossomos compostos por sequências de moléculas chamadas de nucleotídeos, representados pelas letras A, U, G e C, que tem alta capacidade de mutação. Por isso, os vírus se apresentam de formas variadas.

Existem três tipos de vírus influenza que infectam humanos: Influenza A, B e C. Dos três tipos, apenas os dois primeiros estão associados com morbidade e mortalidade sazonal significativa. Os vírus Influenza A, por apresentarem grande variabilidade genética, são divididos em subtipos de acordo com as características antigênicas e genéticas de suas glicoproteínas de superfície, hemaglutinina (HA) e neuraminidase (NA) (Nj e Subbarao, 1999). Existem 16 tipos de hemaglutinina (H1-H16) e 9 tipos de neuraminidase (N1-N9) identificadas em diferentes espécies animais (Gamblin e Skehel, 2010). Atualmente são conhecidas três hemaglutininas (H1, H2 e H3) e duas neuraminidases (N1 e N2) presentes nos vírus influenza do tipo A adaptados para infectar seres humanos, sendo o H1N1 e o H3N2 os mais recorrentes (Latorre-Margalef et al., 2014). O vírus Influenza B é dividido em duas linhagens antigenicamente distintas, a Victoria e a Yamagata (Caini et al., 2018).

O sucesso epidemiológico a longo prazo dos vírus da gripe se deve principalmente ao acumulo de mudanças em seu material genético. Esse processo é chamado de variação antigênica, que ocorre nas duas glicoproteínas de superfície do vírus, o HA e o NA. A variação antigênica torna um indivíduo suscetível a novas cepas, apesar da infecção prévia por vírus influenza ou vacinação prévia, pois o organismo não está preparado para se defender do novo vírus. Essa variação nos vírus influenza A e B é causada pelo acúmulo de mutações nos genes HA e NA (*drift* e *shift* antigênico) (Treanor, 2004).

O *drift* antigênico decorre de alterações genéticas como substituições, deleções e inserções pontuais de nucleotídeos durante a replicação do genoma viral, resultando

na variação genética gradual do vírus (Nj e Subbarao, 1999). Já no *shift* antigênico, que até o momento ocorreu somente com os vírus influenza A, a variação genética é maior, pois o material genético do vírus é reordenado ou misturado. Normalmente o processo *shitf* ocorre através da troca de material genético entre vírus oriundos de diferentes hospedeiros, resultando em um novo subtipo de vírus. Consequentemente, por serem imunologicamente distintos dos vírus anteriores, os novos subtipos podem causar pandemias de gripe (Treanor, 2004). Neste trabalho avaliamos o efeito de *drift*.

2.2 A gripe

A gripe é uma doença infecciosa muito comum causada pelo vírus Influenza afetando aves e mamíferos. A doença normalmente é transmitida por via aérea e por contato direto com superfícies contaminadas. Estima-se que, todos os anos, em torno de 10% da população mundial contraia o vírus, resultando em cerca de 3 a 5 milhões de casos com complicações mais graves e entre 290 mil a 650 mil mortes (Paiva e Toniolo-Neto, 2003; Taubenberger e Morens, 2008).

Os sintomas mais comuns são calafrios, febre, dores de garganta, musculares e de cabeça, tosse, fadiga e sensação geral de desconforto (Eccles, 2005). Entretanto, manifestações mais graves, como insuficiência respiratória e morte, podem ocorrer principalmente com pacientes pertencentes aos grupos de risco. Se encaixam nestes grupos pessoas com mais de 50 anos, adultos e crianças com problemas pulmonares, cardiovasculares, disfunção renal, doenças metabólicas e imunossupressoras, crianças (6 meses a 18 anos) que fazem tratamento com uso prolongado de aspirina e grávidas.

Epidemias de gripe com gravidade variável têm acontecido de maneira sistemática a cada 1 a 3 anos. A maior parte dos casos acontecem em países de clima temperado, principalmente no inverno. No Hemisfério Sul, os surtos ocorrem entre os meses de maio e setembro e no Hemisfério Norte, ocorrem entre novembro e março. Isto porque as temperaturas em torno de 5 a 20°C e umidade relativa próxima de 35% podem favorecer a permanência do vírus em suspensão no ambiente, favorecendo sua propagação (Lowen et al., 2007). Já em países com clima tropical, os períodos de epidemia de influenza são menos distintos e significativos. Em vários países tropicais, um padrão semestral tem sido relatado, com epidemias ocorrendo tanto na primavera quanto no outono, entre as épocas de epidemias de influenza em zonas temperadas (Simonsen, 1999).

Pandemias são mais raras, porém causam danos maiores. A diferença entre epidemia e pandemia é o impacto na população. Enquanto epidemias atingem grande número de pessoas em determinada localidade, as pandemias são amplamente disseminadas pelo mundo. Do início do século XX até o momento, ocorreram quatro pandemias de Influenza nos anos de 1918, 1957, 1968 e 2009. Na primeira delas, cerca de 40 milhões de pessoas ao redor do mundo morreram, e na última, graças a avanços da medicina e da tecnologia o número foi menor, mas ainda significativo, aproximadamente 300 mil (Gill et al., 2010; Dawood et al., 2012).

2.3 Diversidade genética

A diversidade genética é uma medida de biodiversidade. Uma das formas que ela pode ser definida é pela quantidade de diferenças existentes entre sequências de DNA de um grupo de indivíduos. É gerada principalmente por fenômenos de recombinações e mutações genéticas, no caso dos vírus influenza acontecem durante o processo de *drift* e *shift* antigênico. Medir a diversidade genética de uma população tem como objetivo medir toda variação biológica hereditária acumulada durante o processo evolutivo. Essa diversidade genética é gerada, fundamentalmente, por mutação na sequência nucleotídica durante a replicação do RNA (Santos et al., 2015).

Segundo Templeton (2006), o sucesso epidemiológico do vírus influenza ao longo do tempo se dá por sua grande diversidade genética. Essa diversidade é de fundamental importância para que populações de seres vivos se adaptem ao meio ambiente. Quanto maior a diversidade genética, mais adaptável está a espécie para resistir às mudanças ambientais. Dessa maneira, o vírus influenza se torna muito resistente e adaptável.

Quantificar a diversidade genética é uma estratégia para entender a natureza e distribuição de populações (Hamrick et al., 1994). Algumas medidas de diversidade genética são definidas a seguir (Cousins et al., 2012).

A medida % Complexidade é dada por

$$C = \left(\frac{\delta}{N}\right) \times 100, \tag{2.1}$$

onde δ é o número de leituras únicas de nucleotídeos entre sequências e N é o total de leituras de nucleotídeos entre sequências.

Já a medida de Entropia de Shanon, no contexto de diversidade genética, é definida como

$$S = -\left(\frac{1}{\log N}\right) \sum_{i=1}^{n} p_i \log p_i, \tag{2.2}$$

onde p_i é a proporção de leituras com um único padrão de sequências, N é o número total de sequências e n é o número de sequências distintas.

Uma outra medida de diversidade genética muito usual é a elaborada por Nei e Li (1979). A Diversidade de Nei e Li é dada por:

$$D = \sum_{ij} x_i x_j \pi_{ij}, \tag{2.3}$$

onde x_i e x_j são as frequências da i-ésima e j-ésima sequências e π_{ij} é o número médio de diferenças de nucleotídeos por locus (posição da sequência) entre a i-ésima e a j-ésima sequências.

A medida de diversidade de Nei e Li, que foi utilizada neste trabalho, é baseada na heterozigosidade gênica ou alélica. Ao comparar duas sequências genéticas, a medida de Nei e Li (1979) retorna o número médio de diferenças de nucleotídeos entre as duas sequências (Nei e Li, 1979).

3 Metodologia

O objetivo deste estudo é caracterizar através de medidas de diversidade genética e de gráficos a evolução dos vírus influenza H1N1, H3N2, Victoria e Yamagata. Para atingir este objetivo, o trabalho foi realizado em etapas que estão detalhadas neste capítulo, cujos resultados estão apresentados no capítulo 4.

3.1 Montagem do banco de dados

3.1.1 Origem dos dados

Uma forma de compreender a dinâmica dos vírus influenza é através da biologia molecular. A pesquisa a nível molecular proporciona uma forma mais aprofundada de compreender os organismos vivos. A biologia molecular tem aplicação em ampla variedade de problemas que afetam o ser humano, como, por exemplo, prevenção e tratamento de doenças infecciosas, que é o caso da gripe. Pensando nisso, o banco de dados para este estudo foi retirado do site do NCBI (National Center for Biotechnology Information), que é uma instituição que reconhece a importância dos métodos informatizados para a realização de pesquisas biomédicas. Um dos objetivos do NCBI é armazenar conhecimentos sobre biologia molecular, bioquímica e genética e facilitar o uso de tais bancos de dados pela comunidade médica e de pesquisa, deixando eles disponíveis em seu site.

Pelo site `https://www.ncbi.nlm.nih.gov/`, tem-se acesso ao banco de dados do NCBI para os vírus influenza. As informações sobre os vírus são resultado do trabalho de diversos pesquisadores e centros de pesquisa que documentam os casos da doença, sequenciam o vírus e contribuem com seus dados para o banco. Quando o vírus é isolado, sua sequência genética é registrada. Todos os vírus isolados recebem uma nomenclatura que, estabelecida pela Organização Mundial da Saúde em 1980, consiste em: tipo de hospedeiro; região de origem; número da linhagem; ano de isolamento; tipo antigênico das proteínas (H1 a H16) e (N1 a N9) (Memorandum, 1980). Um vírus do banco de dados deste estudo, por exemplo: A/England/456/2009(H1N1) é um Influenza tipo A, isolado primeiramente na Inglaterra, linhagem número 456, ano de 2009 e tipo H1N1.

No site do NCBI, pode-se, através de filtros, selecionar o tipo e o subtipo de vírus, o tipo de sequência, o hospedeiro, a região onde o vírus foi coletado e a proteína que será analisada. Para este trabalho, foram selecionados os vírus Influenza A e B, cujo hospedeiro era humano, que tinham as sequências completas para a proteína HA do cromossomo 4, de todos os continentes e de todos os anos. Como existem muitas

variantes dos vírus Influenza A, foi decidido que apenas os subtipos H1N1 e H3N2 seriam analisados, uma vez que são os mais frequentes do Influenza A. Os vírus do Influenza B, por serem divididos em apenas duas linhagens, foram analisados em sua totalidade. Após filtrar por estas informações, obteve-se um grande banco de dados de 49402 observações com as sequências e informações de data e origem dos vírus H1N1, H3N2 e Influenza B.

Como o objetivo é caracterizar a evolução do vírus ao longo dos anos, os tipos de vírus influenza A e B devem ser analisados separadamente, uma vez que diferem muito entre si. O mesmo se aplica aos subtipos H1N1 e H3N2, do influenza A e as duas linhagens Victoria e Yamagata, do influenza B. As quatro variantes do vírus devem ser separadas para que as análises sejam feitas entre os vírus de uma mesma variante. Dessa forma, o grande banco unificado, com 49404 observações, foi dividido em quatro bancos, um para cada variante do vírus.

Para a montagem do banco de dados do H1N1 e do H3N2 bastou filtrar o conjunto de dados inicial baixado do site pela variável que contém a informação do subtipo do vírus, já que esta informação constava no banco para todos os vírus influenza A. Porém, para a montagem do banco Victoria e do Yamagata, foi necessário fazer uso da análise de agrupamento, já que apenas 12% (1504 de 12444 observações) dos vírus tipo B continham a informação da linhagem a qual pertenciam. Com isso, o banco inicial foi dividido em 3 bancos: banco H1N1, H3N2 e Influenza B. O banco dos vírus H1N1 contém 17795 observações, o do H3N2 contém 19165 e o do Influenza B contém 12444.

3.1.2 Análise de agrupamentos para os vírus Influenza B

Como dito na seção 3.1, apenas 12% (1504 de 12444 observações) dos vírus influenza B estavam classificados por linhagem, Victoria e Yamagata. Assim como os vírus H1N1 e H3N2, do tipo influenza A, devem ser analisados separadamente por serem muito distintos, o mesmo se aplica às linhagens do influenza B. A forma encontrada para solucionar o problema da falta de informação para as linhagens do influenza B foi realizar uma análise de agrupamentos.

Segundo Hair et al. (2009), a análise de agrupamentos é uma técnica analítica para encontrar subgrupos relevantes de indivíduos ou objetos. O objetivo da técnica é classificar uma amostra de entidades (indivíduos ou objetos) em um pequeno número de grupos mutuamente excludentes, com base nas similaridades das entidades. Neste trabalho, o método utilizado para a análise de agrupamento foi o hierárquico, que tem como resultado um dendrograma que indica visualmente a posição de cada entidade hierarquicamente e o grupo em que se encontra. O método hierárquico, apesar de ser mais custoso computacionalmente, foi a forma encontrada para averiguar se de fato havia uma grande separação entre grupos.

Segundo Hair et al. (2009), a medida de similaridade entre objetos é uma medida de semelhança entre os objetos a serem agrupados. Para a análise de agrupamentos, as similaridades entre todos os pares de objetos são previamente calculadas e, a partir disso, a técnica multivariada prossegue agregando objetos semelhantes em agrupamentos. As medidas de similaridade mais usadas são: medidas correlacionais, medidas de distância e medidas de associação. Neste trabalho, a medida de similaridade escolhida para a análise foi a medida de distância, onde valores altos revelam baixa similaridade. O tipo de medida de distância mais comumente utilizado é a

distância euclidiana, que é a distância geométrica no espaço multidimensional, dado por

$$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}, \tag{3.1}$$

onde $\underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$ e $\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix}$ são, respectivamente, os i-ésimos componentes dos vetores x e y, p é o número de dimensões e $d(x,y)$ é a distância entre os objetos ou indivíduos x e y.

Como a distância euclidiana não pode ser diretamente aplicada para este trabalho, a medida de distância usada para calcular a similaridade dos vírus foi a distância genética simples d, que mede o número de diferenças entre as sequências genéticas dos vírus.

Como o intuito de classificar a linhagem de todas as observações do influenza B, a análise de agrupamentos hierárquico foi utilizada com fins de dividir as sequências em dois grandes grupos. Segundo Everitt et al. (2001), quando não há interesse em investigar todos os níveis de hierarquia mas sim dividir os dados em grandes grupos, usa-se a técnica de partição. Em um agrupamento hierárquico a partição é obtida seccionando o dendrograma, que é o resultado gráfico da análise de agrupamentos. Neste caso, o primeiro nível do dendrograma foi seccionado quando foi dividido em dois grupos. Depois de obter a separação das sequências em dois grupos, o objetivo foi validar e identificar qual grupo era da linhagem Yamagata e qual era da Victoria.

Para identificar os dois grupos, fez-se uso dos 12% das sequências que já estavam identificadas no banco de dados. Através da busca dos vírus previamente identificados como Yamagata, 100% (780 sequências) deles estavam dentro de um grupo. Ao buscar os vírus previamente identificados como Victoria, 99,6% (722 de 725 sequências) estavam dentro do outro grupo. Com isso, os vírus do influenza B foram identificados conforme sua linhagem e foi possível transformar o banco de dados único baixado do site NCBI em quatro bancos de dados.

O banco do H1N1 tem o total de 17795 observações, o H3N2 19165, o Victoria 5092 e o Yamagata 7352. Dentre as informações mais importantes que os bancos armazenam estão: informações da data de coleta do material genético, o local de origem e a sequência de DNA de cada vírus. Como as sequências genéticas serão a fonte de informação para as análises deste trabalho, elas tiveram que passar pelo processo de alinhamento, explicado a seguir.

3.1.3 Alinhamento das sequências de DNA

Como o material genético do vírus é replicado em cada geração do vírus, a sequência genética que constitui esse material muda através do processo de mutação. A mutação que acontece na sequência pode ser resultado da substituição, inserção ou deleção de nucleotídeos. Essas mutações no vírus influenza são consequências, principalmente, do processo de *drift* aleatório. Quando essas mudanças acontecem, as sequências genéticas dos vírus começam a divergir, como no exemplo apresentado a seguir, retirado do livro Statistical Method in Bioinformatics, Ewens e Grant (2006):

Sequência original:

$$CGGTATGCCA$$

Sequências descendentes da original:

$$CGGGTATCCAA$$

$$CCCTAGGTCCCA$$

Essa divergência entre as sequências ocorre em taxas e locais variados, dependendo da função que a parte avaliada do DNA executa e como ela tolera mudanças. Muitos problemas da bioinformática são relacionados com a comparação entre duas ou mais sequências de DNA. Para que essa comparação possa ser realizada, usa-se o processo de alinhamento. As duas sequências descendentes acima, depois de passadas pelo processo de alinhamento, podem ser comparadas entre si, uma vez que os nucleotídeos de uma sequência serão comparados em posições equivalentes com nucleotídeos de outra sequência.

Sequências descendentes da original depois do processo de alinhamento:

$$CGGGTA--TCCAA$$

$$CCC-TAGGTCCCA$$

O símbolo "-" é chamado de "indel", que representa a inserção ou a deleção de um nucleotídeo em algum ponto da evolução das sequências. Uma repetição de L "indels" consecutivos é chamada de "gap" de tamanho L. No exemplo acima, a primeira sequência tem um gap de tamanho 2, e a segunda sequência tem um gap de tamanho 1. O alinhamento também permitiu observar que na posição 12 o nucleotídeo, que na sequência original era C, foi substituído pelo A na primeira sequência descendente.

Segundo Ewens e Grant (2006), há vários tipos de métodos de alinhamentos. Existem alinhamentos globais, que alinham a sequência inteira, e existem alinhamentos locais, que alinham apenas subsequências. Existem alinhamentos com gap, que permitem "indels" e alinhamentos sem gap, que não permitem "indels". Há também alinhamento "pairwise", que alinha duas sequências, e o alinhamento múltiplo, usado para alinhar mais de duas sequências.

Dividido o único banco de dados em quatro, um para cada tipo de vírus, as sequências genéticas para cada observação tiveram que ser alinhadas. Como o objetivo do trabalho é descrever a evolução dos Influenza H1N1, H3N2, Victoria e Yamagata, o alinhamento escolhido foi o alinhamento global múltiplo com gap. Este tipo de alinhamento é a generalização do alinhamento "pairwise". Uma forma de escolher o melhor alinhamento de sequências é pelo esquema de escore. Um exemplo de escore é dado por,

$$escore = \left(\frac{m}{T-m}\right), \tag{3.2}$$

onde m é o número de *matches* de nucleotídeos entre todas as sequências e T é o número total de nucleotídeos comparados entre todas as sequências.

Usando essa fórmula, todos os nucleotídeos de uma sequência são comparados com os nucleotídeos em posições equivalentes das outras sequências, para cada combinação de alinhamento. O alinhamento que obtiver o menor escore é considerado

o melhor, pois tem número maior de *matches*. Porém, a medida que o tamanho da sequência aumenta, se torna exaustivo ou inviável listar todas as possibilidades de alinhamento para obter seus escores, especialmente quando o alinhamento é feito com múltiplas sequências. Com isso, vem a necessidade de utilizar métodos computacionais que encontre o melhor escore sem listar todas as possibilidades. Neste trabalho, para realizar o alinhamento, fez-se uso da ferramenta online MAFFT, que é um software para alinhamento múltiplo de sequências de aminoácidos ou nucleotídeos (Yamada et al., 2016). A ferramenta foi desenvolvida para obter resultados mais rápidos para o alinhamento múltiplo e para isso utiliza da transformadas rápidas de Fourier (Ewens e Grant, 2006).

O número de sequências alinhadas para os vírus H1N1, H3N2, Victoria e Yamagata foram, respectivamente, 17795, 19165, 5092, 7352. Os tamanhos das sequências (número de nucleotídeos e indels) para cada tipo de vírus após o alinhamento foram, respectivamente, 2095, 1967, 1913 e 1913. Com as sequências alinhadas substituindo as sequências não alinhadas no banco de dados, finalizou-se a montagem do banco de dados, dando início às análises.

3.2 Diversidade de Nucleotídeos

Para conhecer as diferença entre vírus, fez-se uso de medidas de diversidade, que quantificam a variedade genética entre e dentro de populações. A medida de diversidade de nucleotídeos de Nei e Li, dada na expressão 2.3, que é baseada no número de diferenças de nucleotídeos entre sequências, foi escolhida para realizar as análises. Para obter essa diversidade, a distância entre todas as sequências foi calculada através da função dist.dna do pacote ape do software R. Como resultado, obteve-se, para as quatro variantes estudadas do vírus, uma matriz de distâncias par a par:

$$\mathbf{D} = \begin{pmatrix} d_{11} & d_{12} & d_{13} & \cdots & d_{1n} \\ d_{21} & d_{22} & d_{23} & \cdots & d_{2n} \\ d_{31} & d_{32} & d_{33} & \cdots & d_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & d_{n3} & \cdots & d_{nn} \end{pmatrix},$$

onde d_{ij} é a distância entre as sequências i e j. A distância d_{ij}=0 quando i=j, dado que uma sequência é idêntica a ela mesma. Quando $i \neq j$, d_{ij} pode assumir valores maiores que zero pois duas sequências distintas podem conter diferenças.

Com a matriz de distâncias, tem-se a medida do quanto cada uma das sequências difere das demais. Estas medidas, quando combinadas com as datas de isolamento dos vírus, informam o seu comportamento temporal. A partir disso, pode-se obter medidas de diversidade em um determinado período de tempo, calculando, por exemplo, as distâncias médias das sequências em um período desejado.

Neste estudo, a matriz de distâncias embasou todas as análises. Foram calculadas as distâncias médias por ano para entender e comparar o comportamento das diversidades dos tipos de vírus ao longo do tempo. As diversidades por ano foram obtidas a partir do método

$$D_t = \binom{N_t}{2}^{-1} \sum_{\substack{i<j \\ i,j \in S_t}} d_{ij}, \tag{3.3}$$

onde D_t é a diversidade das sequências no período de tempo t, d_{ij} são as distâncias entre a i-iésima e j-ésima sequências, S_t é o conjunto de índices das sequências com tempos em t, e N_t o número de elementos nesse conjunto S_t.

As diversidades por ano representam a variedade das sequências do ano especificado, ou seja, o quanto, em média, o material dos vírus se diferenciam entre si dentro daquele período de tempo. Foram calculadas as diversidades por ano para as quatro variações de vírus estudada: H1N1, H3N2, Victoria e Yamagata.

3.3 Série temporal em janela deslizante das diversidades

Os resultados das diversidades podem ser ainda mais informativos quando tratados como uma série temporal. Uma característica importante de uma série temporal é que a ordem dos dados é fundamental e as observações vizinhas são dependentes. No caso da evolução do vírus Influenza, dado que as mutações genéticas são cumulativas, é interessante realizar uma análise que considera o ordenamento temporal das observações.

Para investigar mais detalhadamente o comportamento dos vírus ao longo do tempo, as diversidades foram calculadas como uma série temporal no esquema de janela deslizante. Uma série temporal em janela deslizante é uma série de dados ordenados no tempo de comprimento N e analisadas em subsequências de tamanho k. Todas as subsequências possíveis são extraídas pela janela deslizante (Yu et al., 2014), como mostra a imagem 3.1.

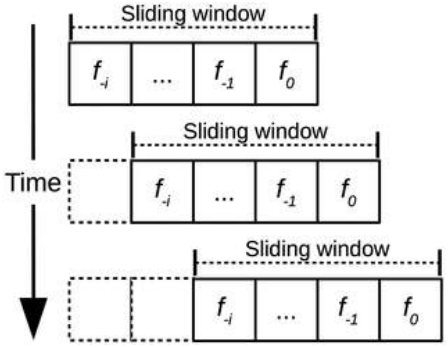

Figura 3.1: Exemplo de janela deslizante. Reproduzida de https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174959

O cálculo da diversidade em série temporal mensal utilizando o esquema de janela deslizante foi feito com três tamanhos de janela $k = 1$, 3 e 6, representando, respectivamente, os períodos mensal, trimestral e semestral. A janela de tamanho 6 calcula a diversidade no tempo t considerando as sequências dos vírus isolados desde o tempo $t-5$ até t. Para a janela de tamanho k=3, são considerados os vírus desde o tempo $t-2$ até t e para a janela de tamanho 1, apenas os vírus no tempo t são considerados. As séries foram calculadas através da fórmula apresentada a seguir.

$$D_{t,k} = \binom{N_{t,k}}{2}^{-1} \sum_{\substack{i<j \\ i,j \in S_{t,k}}} d_{ij}, \tag{3.4}$$

onde $D_{t,k}$ é a diversidade das sequências no período de tempo t considerando os vírus entre os tempos $t-k-1$ até t, d_{ij} são as distâncias entre a i-iésima e j-ésima sequências, $S_{t,k}$ é o conjunto de índices das sequências com tempos entre $t-k-1$ e t, e $N_{t,k}$ o número de elementos nesse conjunto $S_{t,k}$.

Para obter as séries temporais, foram considerados os vírus isolados a partir do ano de 2000 dado que antes deste ano as observações são escassas e com descontinuidades. O tempo $t = 1$, em todos os quatro bancos de dados, representa o mês de janeiro de 2000, o tempo $t = 2$ representa fevereiro de 2000 e assim por diante. Obtidas as séries temporais, uma análise comparativa visual entre os três tamanhos de janela k foi realizada para que fosse definido o melhor período de tempo para calcular as diversidades.

3.4 Visualização gráfica das distâncias

Um dos objetivos deste trabalho é caracterizar visualmente a evolução dos vírus H1N1, H3N2, Victoria e Yamagata. Para alcançá-lo, depois da obtenção dos bancos de dados, uma medida de distância genética foi escolhida para mensurar a diversidade entre sequências. A partir das distâncias entre sequências, pode-se medir as diversidades entre os vírus em determinado período de tempo, por exemplo por ano, semestre, trimestre. Outra forma de utilizar essas distâncias é agrupando-as por similaridade.

Para observar essas similaridades de maneira gráfica, a técnica utilizada foi o escalonamento multidimensional. Segundo (Hair et al., 2009), o objetivo do escalonamento multidimensional é transformar similaridades entre objetos em distâncias representadas em um espaço multidimensional, geralmente de baixa dimensionalidade. Existem três tipos de escalonamento multidimensional: o clássico, o métrico e o não métrico.

Neste trabalho, o tipo de escalonamento multidimensional utilizado foi o clássico, o qual necessita de uma matriz de dessemelhanças entre os pares de itens e produz uma matriz de coordenadas que minimize a função perda, representada a seguir

$$P_D = \left(\frac{\sum_{i,j} (b_{ij} - (x_i, x_j))^2}{\sum_{i,j} b_{ij}^2} \right), \tag{3.5}$$

onde x_i e x_j são as coordenadas que minimizam a função perda P_D e b_{ij} são os m-dimensionais termos da matriz B.

Segundo Wickelmaier (2003), o algoritmo clássico do MDS baseia-se no fato de que minimiza a função perda P_D e X pode ser derivada por decomposição de autovalores a partir da matriz do produto escalar $B = XX'$, que é obtida através dos seguintes passos.

1. Configurar a matriz de distâncias ao quadrado $D^2 = [d_{ij}^2]$.
2. Aplicar a centragem dupla $B = -\frac{1}{2} J P^2 J$, usando $J = I - n^{-1} 11'$, onde n é o número de objetos.

3. Extrair os m maiores autovalores positivos $\lambda_1, \cdots, \lambda_m$ de B e seus respectivos autovetores $e_1, \cdots, e_m$.

4. Uma configuração espacial m-dimensional dos n objetos é derivada da matriz de coordenadas $X = V_m \Lambda_m^{\frac{1}{2}}$, onde V_m é a matriz de m autovetores e Λ_m é matriz diagonal de m autovalores de B, respectivamente.

Se dois vírus com sequências A e B são as mais semelhantes entre si comparados com todos os outros vírus, a técnica de escalonamento multidimensional posicionará graficamente A e B de tal forma que a distância entre elas no espaço multidimensional seja menor do que a distância entre quaisquer outros pares de vírus. Dessa forma, a matriz de distâncias para cada um dos tipos de Influenza estudados foi submetida ao escalonamento multidimensional clássico.

Depois de encontradas as coordenadas x e y, que foram obtidas através da função *cmdscale* do software R, para cada vírus, estas foram representadas graficamente em um plano bidimensional. Para entender melhor o comportamento do vírus, cada uma das sequências posicionadas no gráfico foi colorida conforme o seu ano de isolamento. Dessa forma, pode-se investigar como os dados se distribuem e se esse agrupamento tem relação com o tempo.

4 Resultados

Neste capítulo, são apresentados os resultados das análises descritas no capítulo 3 e suas devidas interpretações.

4.1 O banco de dados

O primeiro passo para a realização deste trabalho foi montar o banco de dados. Ao escolher o site *NCBI*, que contém informações de vírus Influenza e suas respectivas sequências de DNA, as variáveis selecionadas para compor o banco de dados foram:

Acesso - cada vírus, quando é isolado, recebe um código de acesso único.

Tamanho da sequência - número de nucleotídeos no DNA do vírus.

Hospedeiro - a espécie que foi infectada pelo vírus.

Segmento - parte do DNA do vírus avaliada.

Subtipo - subtipo/linhagem do vírus.

Continente - continente em que o vírus foi isolado.

País - país em que o vírus foi isolado.

Data - data em que o vírus foi isolado.

Nome - nome do vírus, obtido através do processo explicado na seção 3.1.1

Completo - se a sequência de DNA do vírus está completa ou não.

Sequência - a sequência de DNA do vírus isolado.

Como decidido previamente, apenas os vírus Influenza A, H1N1 e H3N2, e os Influenza B seriam estudados. Também seriam considerados apenas os vírus que estivessem com a sequência de DNA completa e o segmento a ser investigado seria o 4(HA), pois esse segmento é o mais relevante para estudos evolutivos. Sendo assim, as devidas variáveis foram filtradas.

Depois de baixado o banco de dados, a análise de agrupamento hierárquico foi realizada para classificar as linhagens dos vírus Influenza B, explicado na seção 3.1.2. Antes da análise, 88% das observações da variável *Subtipo* eram dados faltantes.

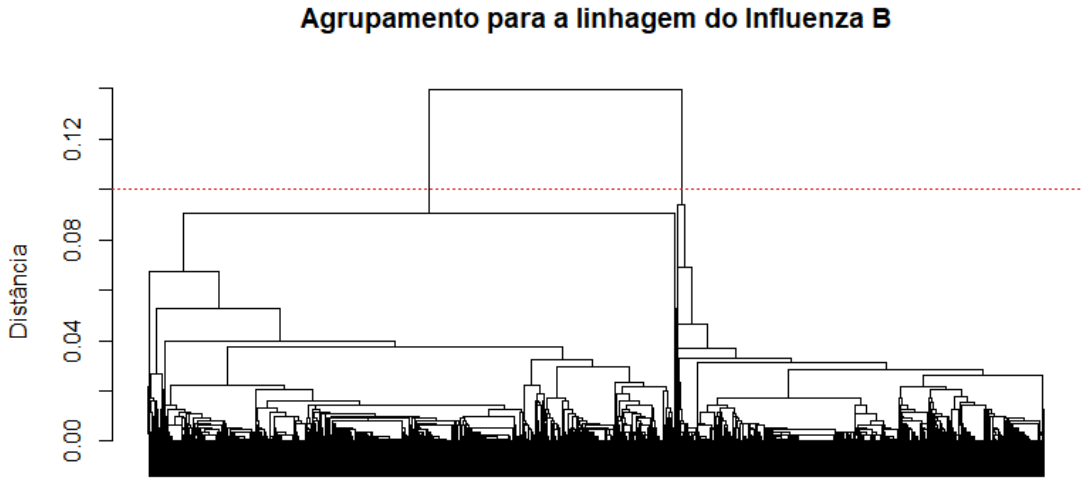

Figura 4.1: dendrograma da linhagem do Influenza B

O dendrograma da figura 4.1 obtido com o método de agrupamento hierárquico, foi seccionado quando se dividiu em dois grupos. Isso porque a intenção desta análise era separar os vírus em duas linhagens. Todas os vírus ligados ao mesmo grande grupo foram considerados da mesma linhagem. Depois, foi feita a busca dos 12% dos vírus dentro do dendrograma que estavam classificados previamente. O resultado foi que 100% dos vírus Yamagata estavam dentro do mesmo grupo e 99,6% dos Victoria estavam em outro. Dessa forma, os vírus tipo B foram todos classificados por linhagem, e a variável *Subtipo* foi totalmente preenchida. A análise do dendrograma da figura 4.1 evidencia a grande separação entre os dois grandes grupos e corrobora o uso desta técnica para separar as linhagens.

Depois que a variável *Subtipo* foi devidamente preenchida e a variável *Sequência* passou pelo processo de alinhamento, descrito em 3.1.3, iniciou-se o processo de análise dos dados. As informações do número de observações em cada país para cada tipo de vírus estão representadas nas tabelas abaixo:

Tabela 4.1: Frequência de observações para cada continente no banco de dados

Continente	H1N1	H3N2	Yamagata	Victoria	Total
África	149	102	56	28	335
América do Norte	9166	13526	5412	3995	32099
América do Sul	615	564	18	30	1227
Ásia	4885	2796	885	477	9043
Europa	2558	801	21	1	3381
Oceania	422	1376	960	561	3319
Total	17795	19165	7352	5092	49404

Os vírus Influenza A representam 74,8% das observações do banco de dados,

sendo que 38,8% pertencem ao subtipo H3N2 e 36% ao subtipo H1N1. Os vírus Influenza B representam 25,2% das observações, 14,9% delas pertencem à linhagem Yamagata e 10,3% à Victoria.

O somatório de todos as observações dos quatro banco de dados é 49404 e o total de observações da América do Norte é 39099, representando 65% da totalidade dos dados. A África tem representatividade menor que 1% nos dados, o que dificultaria a análise dos vírus desse continente. A América do Sul e a Europa também apresentam escassez de dados, principalmente se tratando do vírus Influenza B. A Oceania contém menos dados que a Europa mas, por ser o menor dos continente, pode-se dizer que ela está bem representada no banco de dados. A Ásia contribui com 18% dos dados.

Essa distribuição geográfica desigual do número de sequências é um reflexo do volume de pesquisas científicas em cada local, e não corresponde à distribuição global de casos de gripe. Ao analisar os resultados desse trabalho devemos atentar para o fato de que nossas medidas de diversidade estão desproporcionalmente afetadas pela dinâmica da América do Norte.

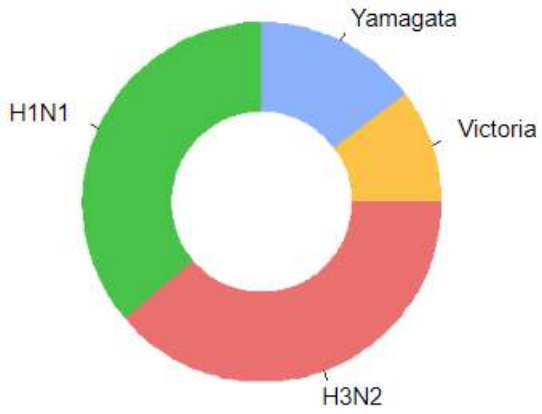

Figura 4.2: Representatividade dos tipos de vírus no banco de dados

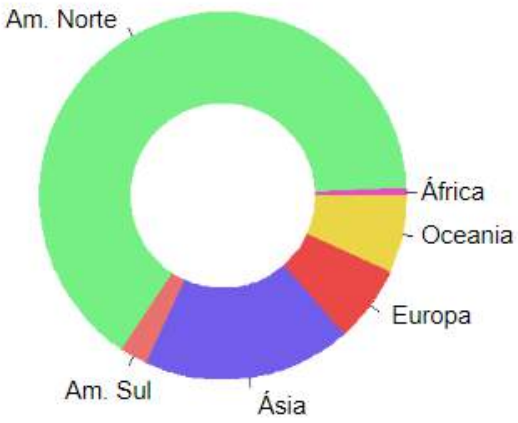

Figura 4.3: Representatividade dos continentes no banco de dados

4.2 Diversidade por ano

As figuras 4.4, 4.5, 4.6 e 4.7 apresentam a evolução da diversidade anual do vírus obtida conforme seção 3.2. As diversidades das sequências por ano, apresentam baixa qualidade nos dados, principalmente antes de 2000 para os vírus H1N1. Para analisar uma série, é interessante que ela tenha continuidade. Percebe-se também que o pico esperado para o gráfico do H1N1 era no ano de 2009, devido à pandemia de H1N1, porém isso não ocorreu.

Por isso, a análise que está na próxima seção foi realizada no esquema de janela deslizante, para que as análises sejam feitas com dados com maior continuidade, e apenas os vírus a partir de 2000 foram analisados.

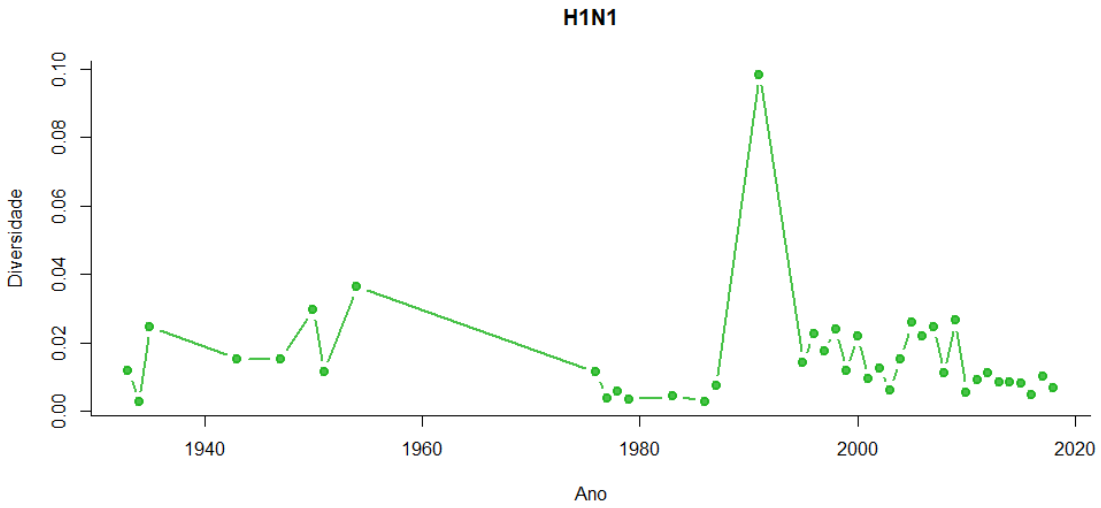

Figura 4.4: Diversidade genética do vírus H1N1 por ano

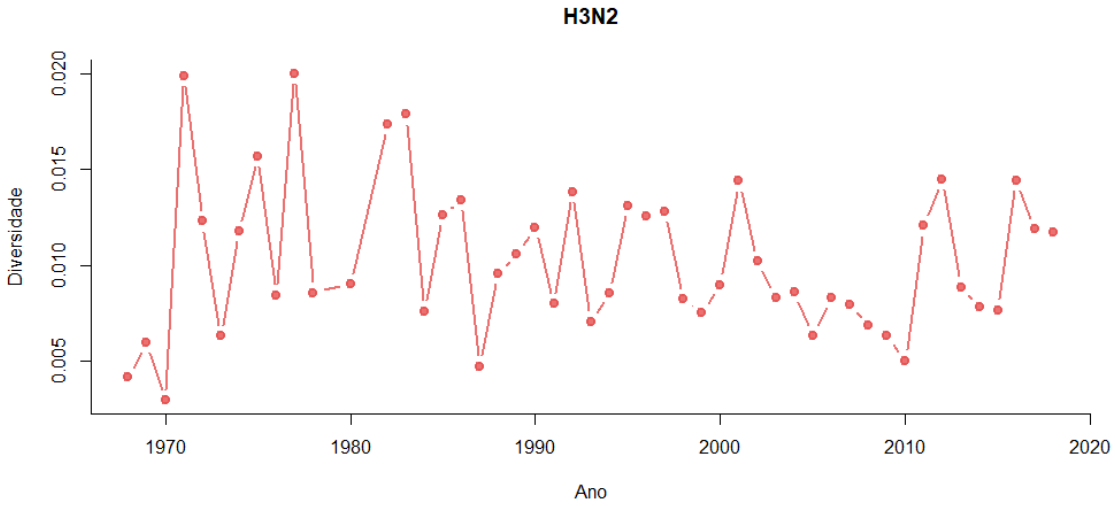

Figura 4.5: Diversidade genética do vírus H3N2 por ano

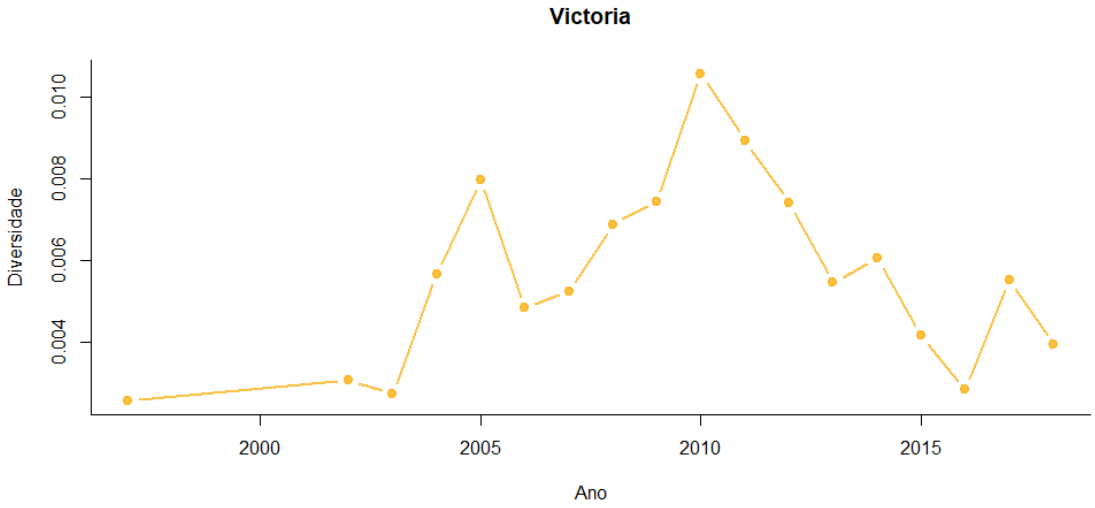

Figura 4.6: Diversidade genética do vírus Victoria por ano

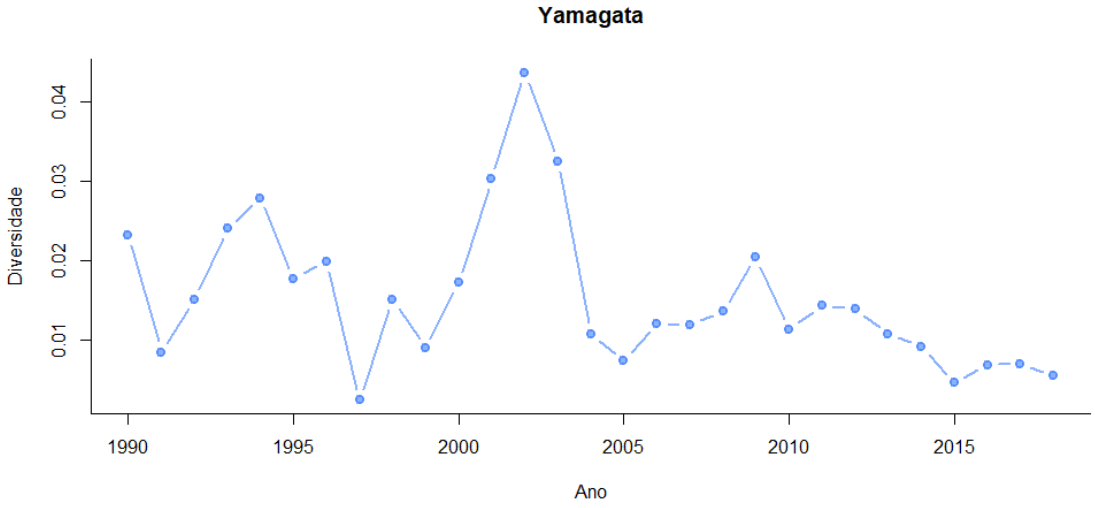

Figura 4.7: Diversidade genética do vírus Yamagata por ano

4.3 Diversidade em série temporal com janela deslizante

4.3.1 Escolha do tamanho da janela

Com o intuito de analisar o comportamento dos vírus com maior precisão e de forma contínua no tempo, as diversidades foram obtidas em formato de série temporal, no esquema de janela deslizante, como foi explicado na seção 3.3.

Para escolher um tamanho de janela que melhor descreva o comportamento dos vírus, os gráficos das diversidades dos vírus H1N1 foram usados como parâmetro para comparar os tamanhos de janelas diferente. Espera-se que quanto maior for o tamanho da janela, mais suave seja o comportamento da série.

As figuras 4.8, 4.9 e 4.10 apresentam estas séries com tamanhos de janela desli-

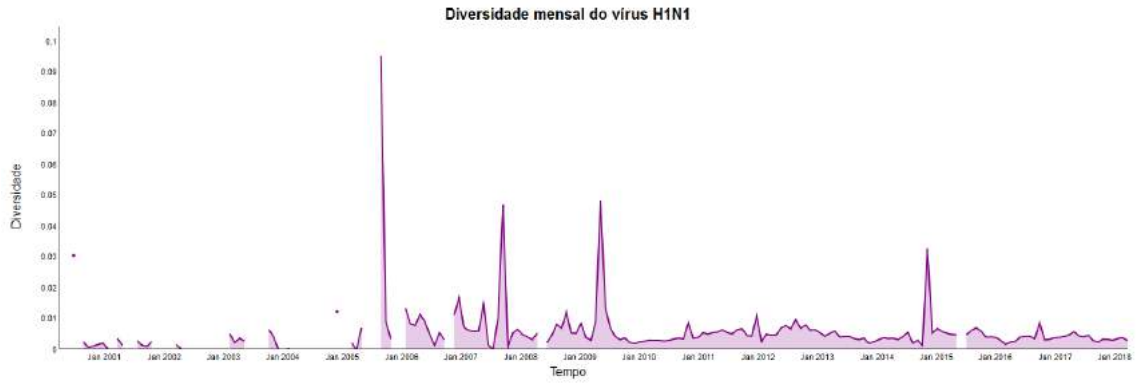

Figura 4.8: Diversidade do vírus H1N1 em série mensal com k=1

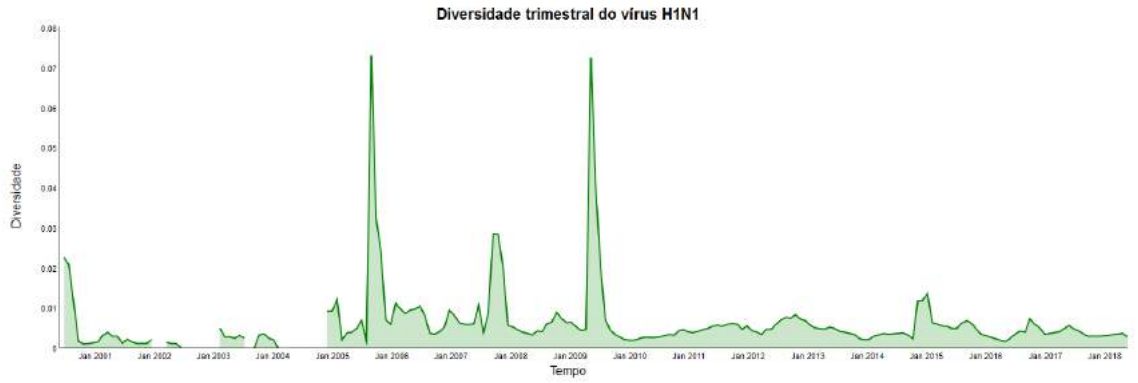

Figura 4.9: Diversidade do vírus H1N1 em série trimestral com k=3

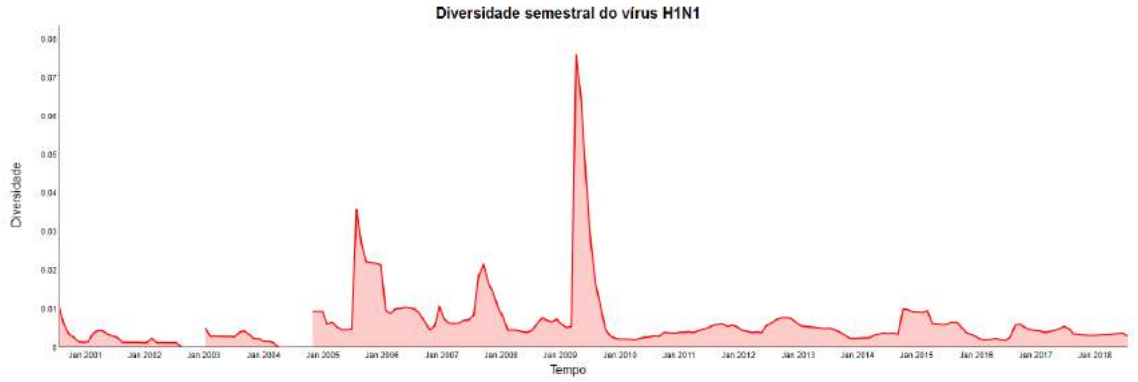

Figura 4.10: Diversidade do vírus H1N1 em série semestral com k=6

zante $k = 1$, 3 e 6, respectivamente. Como esperado, o tamanho da janela impactou na suavidade da série. Como o gráfico das diversidades mensais tem muitas descontinuidades e o gráfico das diversidades semestrais suaviza a série perdendo a informação de picos de diversidade, o tamanho de janela escolhida para avaliar o comportamento dos outros vírus foi $k = 3$, trimestral. As séries com $k = 1$ e 6 para os vírus H3N2, Victoria e Yamagata estão apresentados no apêndice.

4.3.2 Análise das diversidades em série temporal trimestral

As figuras 4.11, 4.12, 4.13 e 4.14 são referentes às diversidades em série trimestral dos vírus Influenza H1N1, H3N2, Victoria e Yamagata.

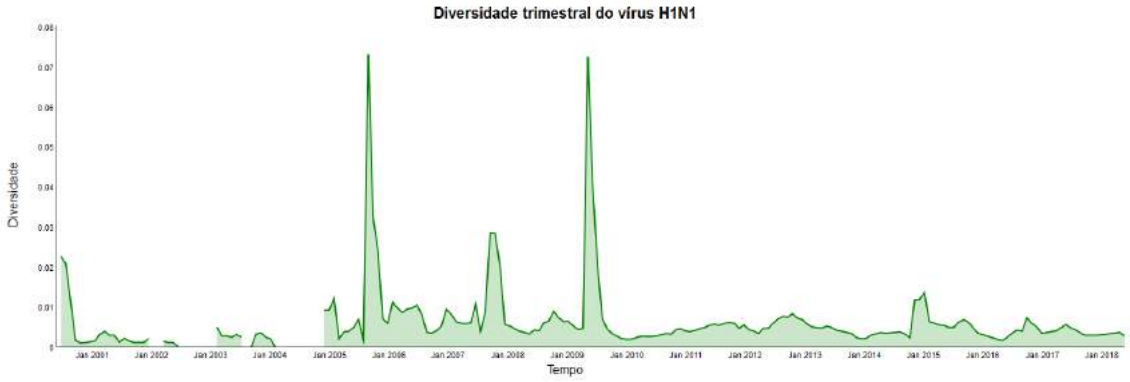

Figura 4.11: Diversidade do vírus H1N1 em série temporal trimestral

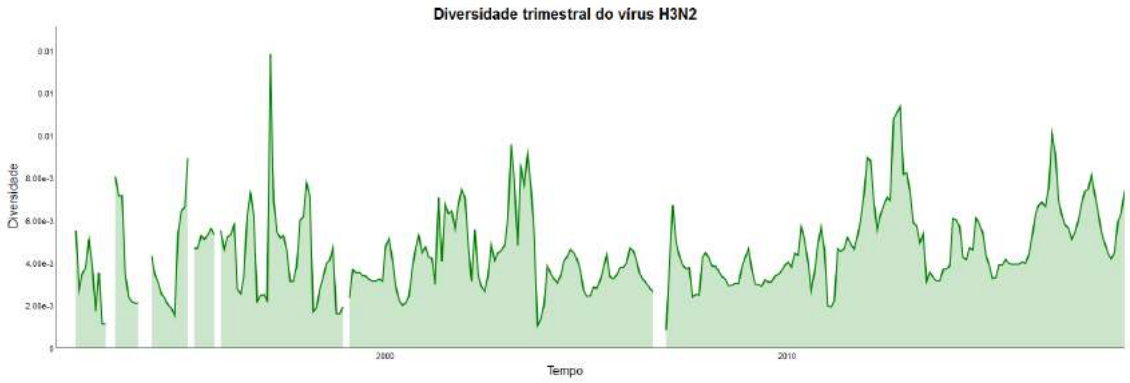

Figura 4.12: Diversidade do vírus H3N2 em série temporal trimestral

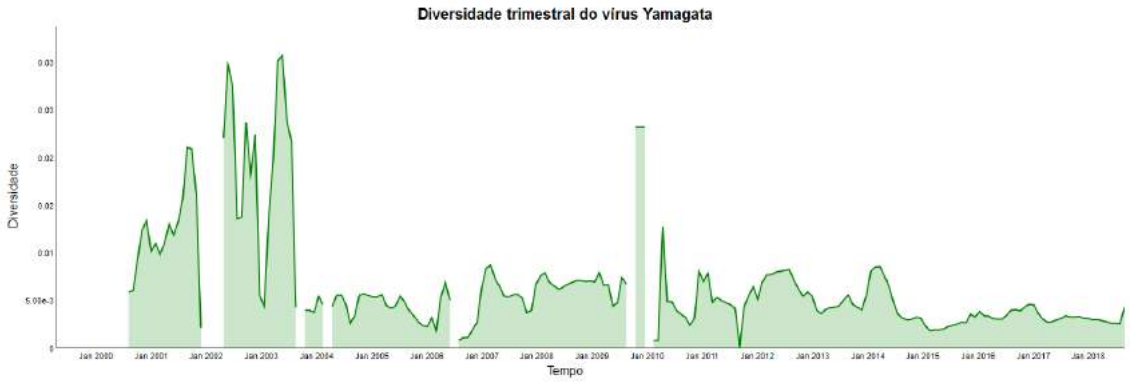

Figura 4.13: Diversidade do vírus Yamagata em série temporal trimestral

O gráfico das diversidades em série temporal com janela deslizante de tamanho $k = 3$ do vírus H1N1 contém algumas descontinuidades entre 2002 e 2004. Em geral, a diversidade do H1N1 tem dimensão parecida com a dos vírus Victoria e Yamagata, ficando entre 0 e 0,03. A exceção são os dois picos nos anos de 2005 e

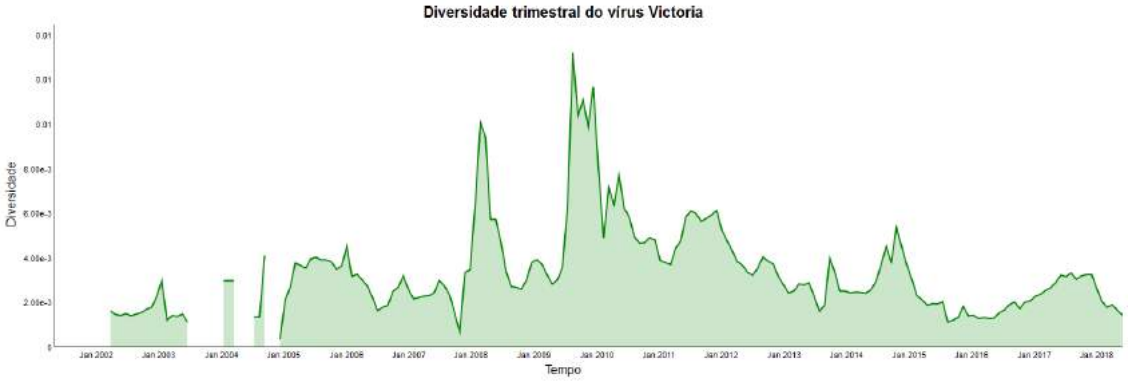

Figura 4.14: Diversidade do vírus Victoria em série temporal trimestral

2009. A alta diversidade genética em 2005 pode ter relação com a amostra pequena (63 observações), já a alta diversidade genética no ano de 2009 coincide com a pândemia do vírus influenza H1N1.

Analisando o gráfico do vírus H3N2, percebe-se que a dimensão da sua diversidade é menor quando comparada com a dos outros vírus e seu comportamento é mais uniforme. O comportamento não parece ter sazonalidade e há uma pequena tendência no aumento das diversidades trimestrais a partir do ano de 2012, podendo ser um indício de que o vírus H3N2 está mais adaptável. O vírus Victoria e Yamagata, do influenza B, tem comportamentos bem semelhantes, com as diversidades ficando entre 0 e 0,03.

4.4 Visualização das distâncias

A partir da técnica multivariada de escalonamento multidimensional, mencionada na seção 3.4, obteve-se os mapas de diversidade genética para cada tipo de vírus. As figuras 4.15, 4.16, 4.17, 4.18, 4.19 e 4.20 são a representação dos vírus em um espaço bidimensional. As coordenadas de cada vírus foram obtidas através da matriz de distâncias genéticas.

Cada ponto no gráfico representa um vírus e cada cor representa um ano. Em todos os gráficos, pode-se perceber que os vírus se agrupam por cores, ou seja, por ano. Esse resultado demonstra que o nível de similaridade entre dois vírus depende de sua distância no tempo. Os vírus de 2005, por exemplo, são mais similares aos vírus de 2004 do que aos de 2000. Isso porque quanto maior o período entre dois vírus, mais sujeitos eles estão às mutações genéticas.

Pode-se perceber também que a figura 4.15 do H1N1 mostra dois grupos bem definidos, o primeiro com os vírus até cerca de 2009 e o segundo com os vírus entre os anos de 2009 e 2018. Isso aconteceu devido à pandemia do H1N1 em 2009, quando ocorreu um surgimento de uma nova linhagem geneticamente distintas à anterior. Por isso, duas análises separadas para o vírus H1N1 foram feitas: o H1N1 pré-pandemia, figura 4.16 e o H1N1 pandêmico, figura 4.17.

Na figura 4.11 da seção 4.3.2, a diversidade dos vírus em 2009 foi muito elevada em relação aos outros anos. Já na figura 4.5 da seção 4.2, onde as diversidades foram calculadas por ano, a alta diversidade do H1N1 em 2009 não foi representada. Nas figuras do H1N1 pré-pandemia e H1N1 pandêmico, percebe-se que existem bandas

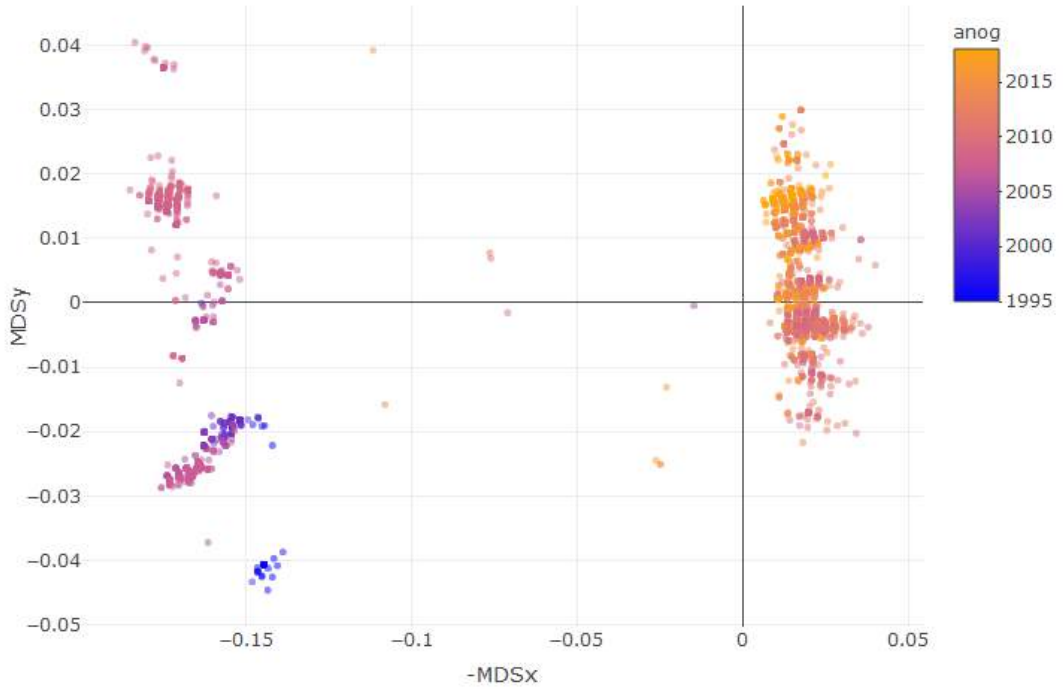

Figura 4.15: Mapa da evolução da diversidade genética para o H1N1.

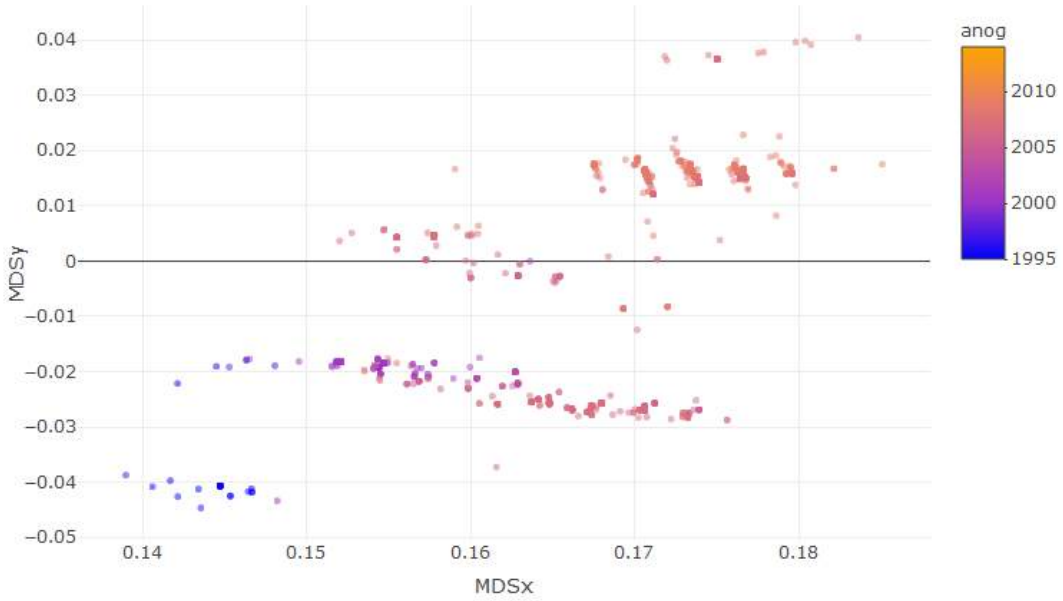

Figura 4.16: Mapa da evolução da diversidade genética para o H1N1 pré-pandemia.

de sequências ao longo do tempo que se agrupam por ano, e essas se direcionam para a direita.

O vírus H3N2, a partir de 2010, visualmente vêm apresentando maior variabilidade, conforme a figura 4.18. Essa característica pode ser atribuída à aparente elevação da diversidade nos últimos anos, já constatada no gráfico das diversidades

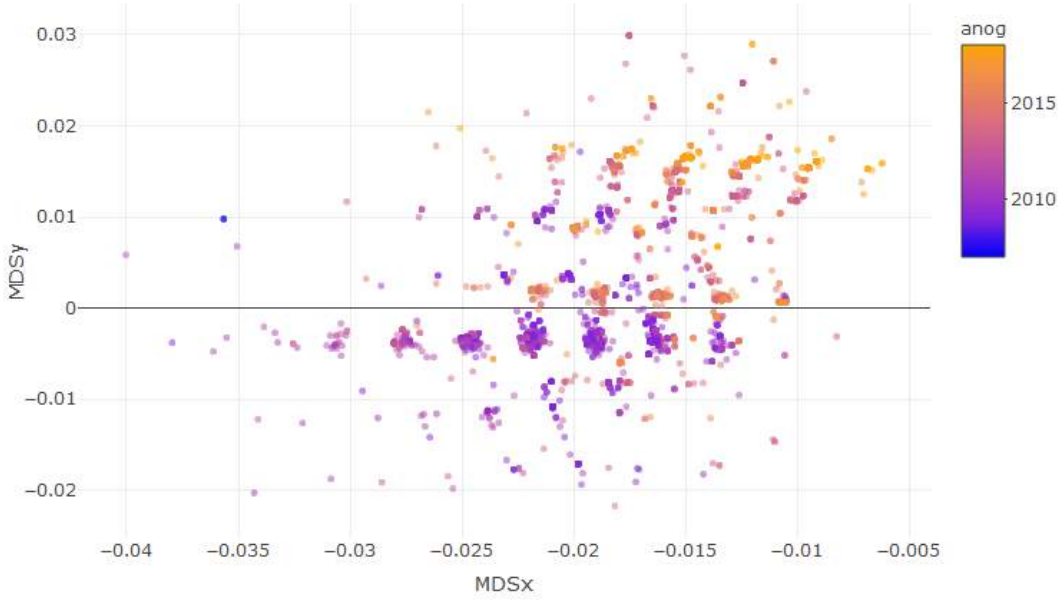

Figura 4.17: Mapa da evolução da diversidade genética para o H1N1 pandêmico.

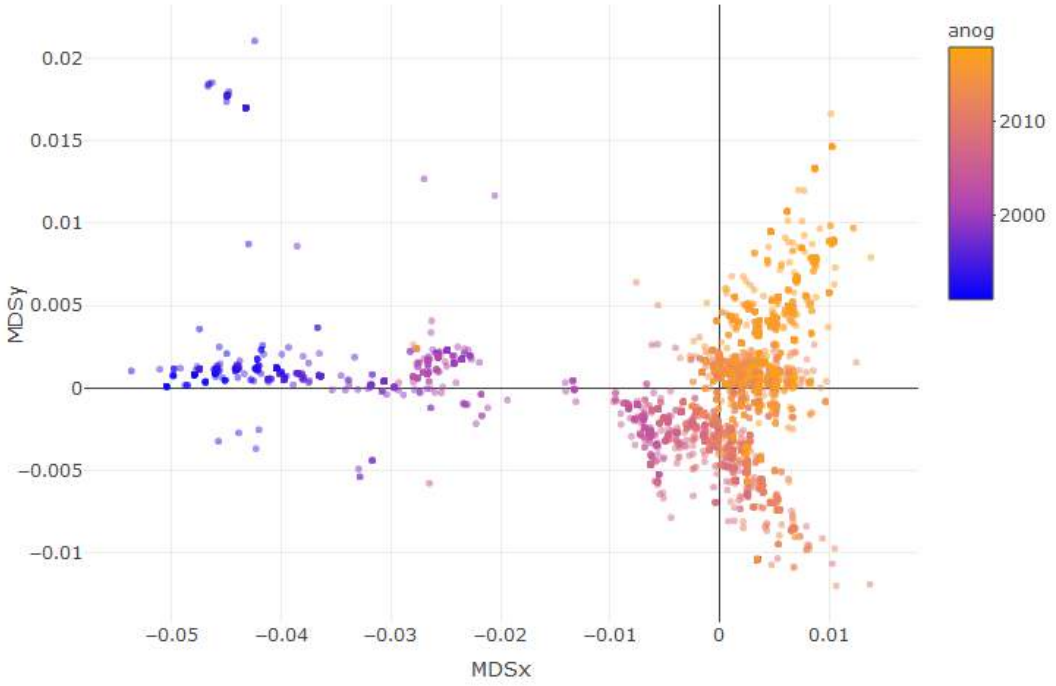

Figura 4.18: Mapa da evolução da diversidade genética para o H3N2.

em série temporal com janela deslizante, na figura 4.12 da seção 4.3.2, mas não percebida no gráfico de diversidade por ano, na figura 4.5 da seção 4.2.

Os vírus Victoria e Yamagata apresentam menor variabilidade entre suas sequências, pois se posicionam mais juntas ao longo do gráfico. Isso acontece principalmente a partir do ano de 2005, conforme os gráficos 4.19 e 4.20. No gráfico da diversidade

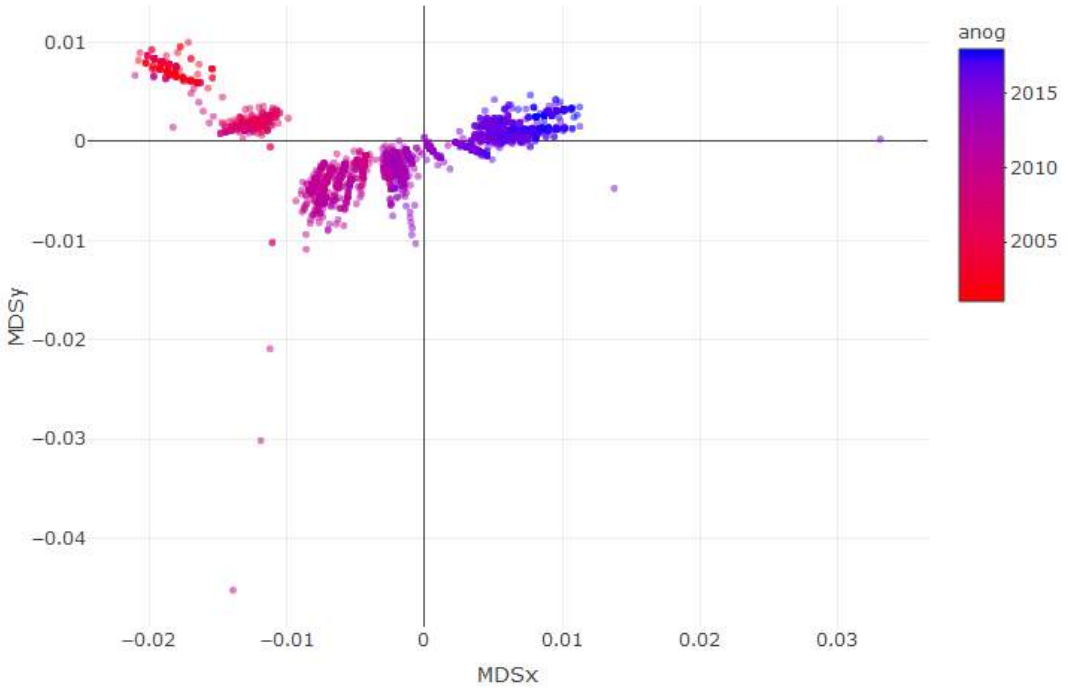

Figura 4.19: Mapa da evolução da diversidade genética para o Victoria.

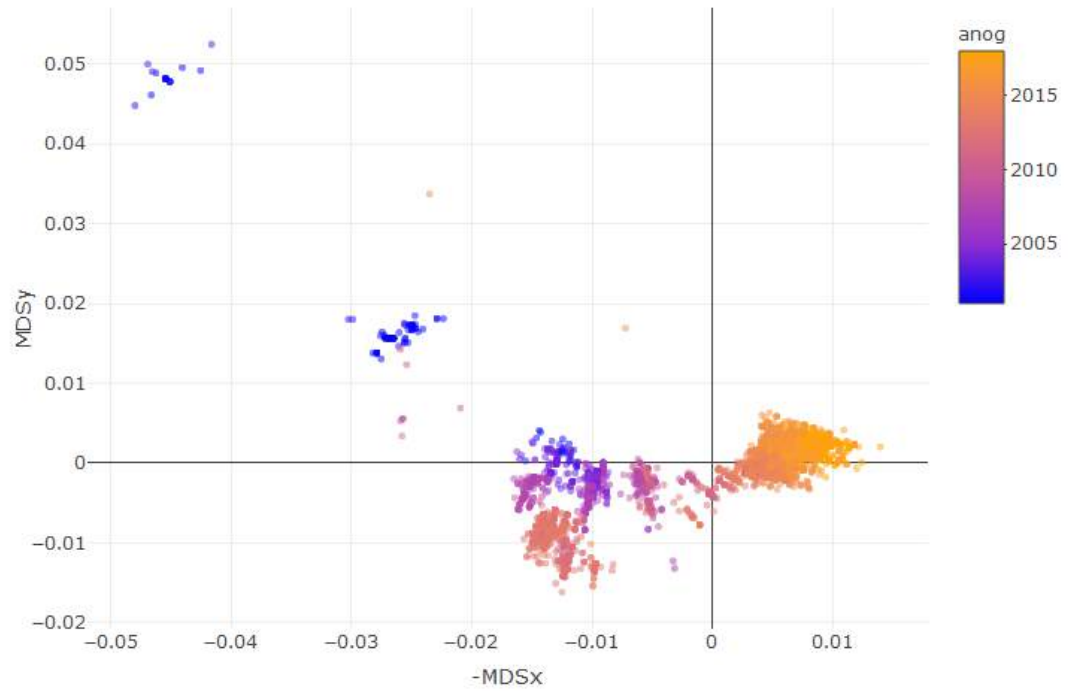

Figura 4.20: Mapa da evolução da diversidade genética para o Yamagata.

trimestral, os dois vírus obtiveram diversidades semelhantes, com comportamento uniforme a partir de 2004. As duas linhagens de vírus antes de 2004, nos gráficos 4.14 e 4.13, apresentam maiores diversidades e maiores variabilidades, o que também

se constata nos gráficos obtidos através do escalonamento multidimensional, onde se percebe um salto na distância entre os vírus anteriores a 2004 e posteriores a este ano.

5 Considerações finais

Este trabalho foi desenvolvido com o intuito de agregar informações sobre o comportamento da diversidade dos vírus Influenza ao longo do tempo. A motivação para o estudo vem de um problema real causado pelo Influenza: a gripe. Ano após ano, a doença afeta milhares de pessoas ao redor do mundo. Isso porque o vírus possui altas taxas de mutação em seu material genético, tornando-o muito adaptável à mudanças ambientais e, dessa forma, permanecendo em circulação ao longo dos anos.

O objetivo desse estudo foi caracterizar a evolução dos vírus Influenza ao longo do tempo a partir da medida de distância genética e da diversidade de Nei e Li (1979) entre sequências. Para que essas medidas fossem melhor compreendidas, fez-se uso de representações gráficas. Antes de iniciar as análises das diversidades dos vírus, as observações do Influenza B foram submetidas à análise de agrupamento hierárquico para que suas linhagens fossem separadas. A partir do dendrograma, resultado desta análise, percebe-se que existem uma diferença grande entre linhagens diferentes dos vírus.

Depois de realizada a separação do único banco de dados em outros quatro bancos, as diversidades foram mensuradas em série temporal por ano. Posteriormente, as diversidades foram obtidas em série temporal no esquema de janela deslizante, que agrega valor para a análise dado que é interessante considerar as diversidades nos tempos anteriores já que os vírus sofrem mutações contínuas em seu material genético. Outra análise realizada fez uso das distâncias entre as sequências para representa-las em um espaço bidimensional, através da técnica de escalonamento multidimensional.

Nos gráficos apresentados neste trabalho, pode-se perceber que os resultados das diversidades em série temporal com janela deslizante condizem com os resultados dos gráficos do escalonamento multidimensional, o que é indicio de que são duas boas formas de representar a variabilidade genética dos vírus. A série temporal com tamanho de janela $k = 3$ é a que melhor representa a diversidade genética dos vírus ao longo do tempo, entre os três tamanhos $k = 1$, 3 e 6, ou seja, monitorar as diversidade por trimestre pode ser uma boa maneira de entender o comportamento dos vírus. Os gráficos das diversidades por ano não foram muito informativos. Isso porque sua janela temporal era maior (12 meses) e o cálculo das diversidades foi feito em intervalos fixos no tempo, sem considerar as diversidades dos vírus em tempos anteriores, perdendo o efeito sazonal dos vírus.

Em futuras contribuições para este assunto, outras medidas de diversidade genética podem ser avaliadas para comparar com os resultados obtidos neste trabalho

utilizando a diversidade de Nei e Li. Como dito na seção 2.3, a alta diversidade dentro de uma espécie torna ela mais adaptável ao meio. Com isso, uma futura pesquisa poderia avaliar se o número de casos de gripe esta associado à diversidade dos vírus.

Referências Bibliográficas

Caini, S., Kroneman, M., Wiegers, T., Guerche Séblain, C. e., e Paget, J. (2018). Clinical characteristics and severity of influenza infections by virus (sub) type: a literature review.

Cousins, M. M., Ou, S.-S., Wawer, M. J., Munshaw, S., Swan, D., Magaret, C. A., Mullis, C. E., Serwadda, D., Porcella, S. F., Gray, R. H., et al. (2012). Comparison of a high resolution melting (hrm) assay to next generation sequencing for analysis of hiv diversity. *Journal of clinical microbiology*, pages JCM–01460.

Dawood, F. S., Iuliano, A. D., Reed, C., Meltzer, M. I., Shay, D. K., Cheng, P.-Y., Bandaranayake, D., Breiman, R. F., Brooks, W. A., Buchy, P., et al. (2012). Estimated global mortality associated with the first 12 months of 2009 pandemic influenza a h1n1 virus circulation: a modelling study. *The Lancet infectious diseases*, 12(9):687–695.

Eccles, R. (2005). Understanding the symptoms of the common cold and influenza. *The Lancet infectious diseases*, 5(11):718–725.

Everitt, B. S., Dunn, G., et al. (2001). *Applied multivariate data analysis*, volume 2. Wiley Online Library.

Ewens, W. J. e Grant, G. R. (2006). *Statistical methods in bioinformatics: an introduction*. Springer Science & Business Media.

Gamblin, S. J. e Skehel, J. J. (2010). Influenza haemagglutinin and neuraminidase membrane glycoproteins. *Journal of Biological Chemistry*, pages jbc–R110.

Gill, J. R., Sheng, Z.-M., Ely, S. F., Guinee Jr, D. G., Beasley, M. B., Suh, J., Deshpande, C., Mollura, D. J., Morens, D. M., Bray, M., et al. (2010). Pulmonary pathologic findings of fatal 2009 pandemic influenza a/h1n1 viral infections. *Archives of pathology & laboratory medicine*, 134(2):235–243.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., e Tatham, R. L. (2009). *Análise multivariada de dados*. Bookman Editora.

Hamrick, J. L., Schnabel, A., e Wells, P. V. (1994). Distribution of genetic diversity within and among populations of great basin conifers. *Natural history of the Colorado plateau and great Basin*, pages 147–161.

Latorre-Margalef, N., Tolf, C., Grosbois, V., Avril, A., Bengtsson, D., Wille, M., Osterhaus, A. D., Fouchier, R. A., Olsen, B., e Waldenström, J. (2014). Long-term variation in influenza a virus prevalence and subtype diversity in migratory mallards in northern europe. *Proc. R. Soc. B*, 281(1781):20140098.

Lowen, A. C., Mubareka, S., Steel, J., e Palese, P. (2007). Influenza virus transmission is dependent on relative humidity and temperature. *PLoS pathogens*, 3(10):e151.

Memorandum, W. (1980). A revision of the system of nomenclature for influenza viruses: a who memorandum. *Bull World Health Organ*, 58:585–91.

Neher, R. A. e Bedford, T. (2015). nextflu: Real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics*, 31(21):3546–3548.

Nei, M. e Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, 76(10):5269–5273.

Nj, C. e Subbarao, K. (1999). Influenza. *Lancet*, 354:1277–82.

Paiva, T. M. e Toniolo-Neto, J. (2003). Eduardo forleo-neto1, elisa halker1, verônica jorge santos1. *Revista da Sociedade Brasileira de Medicina Tropical*, 36(2):267–274.

Santos, F. R., Lacerda, D., Redondo, R., Nascimento, A., Chartone-Souza, E., Borba, E., Ribeiro, R., e Lovato, M. (2015). Diversidade genética. *Biota Vida*, pages 389–410.

Simonsen, L. (1999). The global impact of influenza on morbidity and mortality. *Vaccine*, 17:S3–S10.

Stöhr, K. (2002). Influenza—who cares. *The Lancet infectious diseases*, 2(9):517.

Taubenberger, J. K. e Morens, D. M. (2008). The pathology of influenza virus infections. *Annu. Rev. pathmechdis. Mech. Dis.*, 3:499–522.

Templeton, A. R. (2006). *Population genetics and microevolutionary theory*. John Wiley & Sons.

Treanor, J. (2004). Influenza vaccine—outmaneuvering antigenic shift and drift. *New England Journal of Medicine*, 350(3):218–220.

Wickelmaier, F. (2003). An introduction to mds. *Sound Quality Research Unit, Aalborg University, Denmark*, 46(5).

Yamada, K. D., Tomii, K., e Katoh, K. (2016). Application of the mafft sequence alignment program to large data—reexamination of the usefulness of chained guide trees. *Bioinformatics*, 32(21):3246–3251.

Yu, Y., Zhu, Y., Li, S., e Wan, D. (2014). Time series outlier detection based on sliding window prediction. *Mathematical problems in Engineering*, 2014.

6 Apêndice

Estão apresentados aqui as séries em esquema de janela deslizante para $k = 1$ e 6, para as linhagens H3N2, figuras 6.1 e 6.2, Yamagata, figuras 6.3 e 6.4 e Victoria, figuras 6.5 e 6.6.

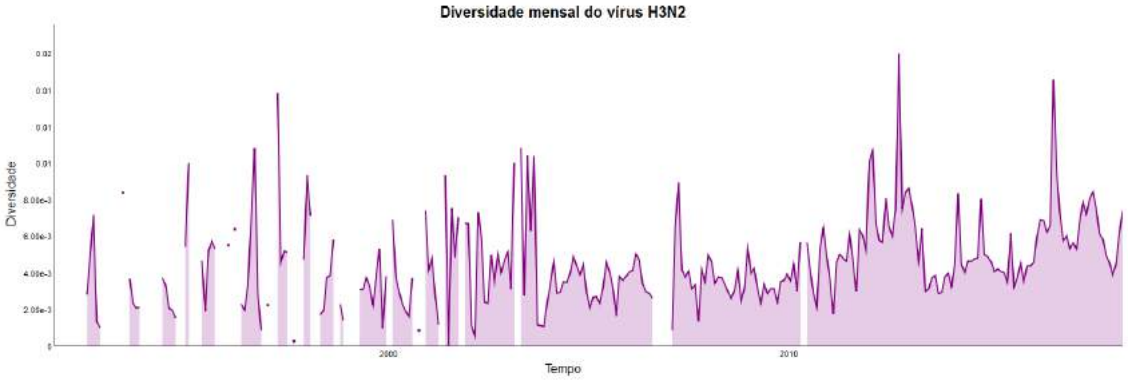

Figura 6.1: Diversidade do vírus H3N2 em série temporal mensal

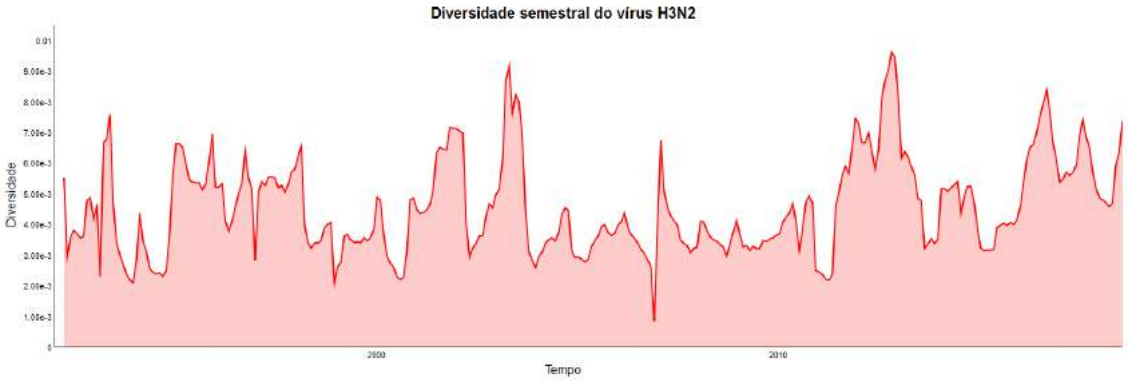

Figura 6.2: Diversidade do vírus H3N2 em série temporal semestral

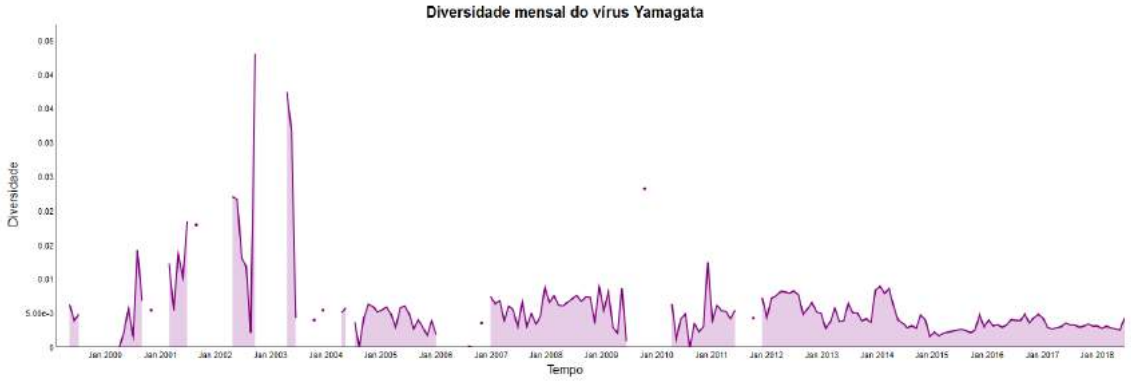

Figura 6.3: Diversidade do vírus Yamagata em série temporal mensal

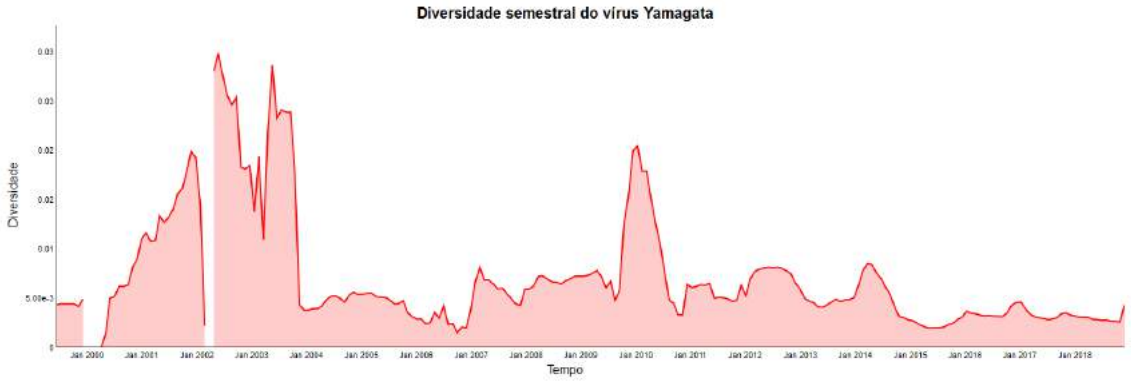

Figura 6.4: Diversidade do vírus Yamagata em série temporal semestral

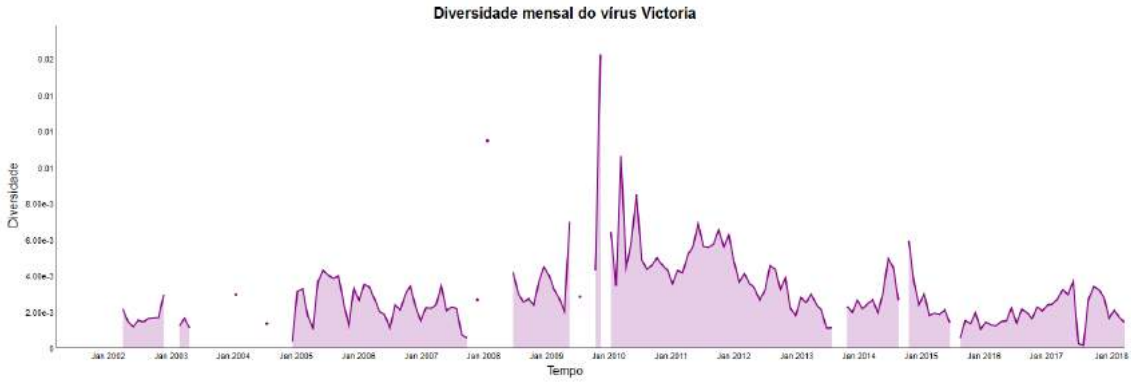

Figura 6.5: Diversidade do vírus Victoria em série temporal mensal

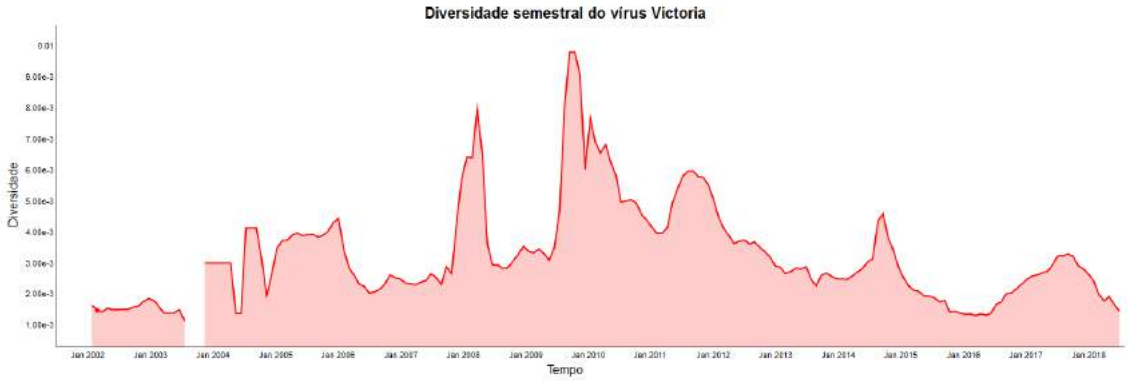

Figura 6.6: Diversidade do vírus Victoria em série temporal semestral