

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Álvaro Salgado de Abreu

**APRENDIZADO DE MÁQUINA APLICADO À VIGILÂNCIA GENÔMICA DE VÍRUS
EMERGENTES E REEMERGENTES**

Belo Horizonte

2022

Álvaro Salgado de Abreu

**APRENDIZADO DE MÁQUINA APLICADO À VIGILÂNCIA GENÔMICA DE VÍRUS
EMERGENTES E REEMERGENTES**

Versão final

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais, como requisito para obtenção do grau de Doutor em Bioinformática

Orientador: Prof. Dr. Luiz Carlos Júnior Alcantara

Coorientador: Prof. Dr. José Lourenço

Belo Horizonte

2022

043

Abreu, Álvaro Salgado de.

Aprendizado de Máquina aplicado à vigilância genômica de vírus emergentes e reemergentes [manuscrito] / Álvaro Salgado de Abreu. – 2022. 78 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Luiz Carlos Júnior Alcantara. Coorientador: Prof. Dr. José Lourenço.

Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Aprendizado de Máquina. 3. Genômica. 4. Inteligência Artificial. 5. Vírus da Febre Amarela. 6. Vírus da Dengue. I. Alcantara, Luiz Carlos Júnior. II. Lourenço, Jose Miguel Lima de Aguiar Graça. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004



UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

ATA DE DEFESA DE TESE

ÁLVARO SALGADO DE ABREU

Às treze horas do dia **14 de junho de 2022**, reuniu-se, por videoconferência, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: "**Aprendizado de Máquina aplicado à vigilância genômica de vírus emergentes e reemergentes**", requisito para obtenção do grau de Doutor em **Bioinformática**. Abrindo a sessão, o Presidente da Comissão, **Dr. Luiz Carlos Júnior Alcântara**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Professor/Pesquisador	Instituição	Indicação
Dr. Luiz Carlos Júnior Alcântara	FIOCRUZ	Aprovado
Dr. José Lourenço	University Of Oxford	Aprovado
Dr. Aristóteles Góes Neto	Universidade Federal de Minas Gerais	Aprovado
Dr. Dennis Maletich Junqueira	Universidade Federal do Rio Grande do Sul	Aprovado
Dr. Diego Gervasio Frías Suárez	UNEB	Aprovado
Dr. Fabrício Murai Ferreira	Universidade Federal de Minas Gerais	Aprovado

Pelas indicações, o candidato foi considerado: **Aprovado**

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 14 de junho de 2022.



Documento assinado eletronicamente por **Diego Gervasio Frías Suarez, Usuário Externo**, em 14/06/2022, às 16:56, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).

Documento assinado eletronicamente por **Dennis Maletich Junqueira, Usuário Externo**, em 15/06/2022, às

29/07/2022 13:57

SEI/UFMG - 1495622 - Ata



08:20, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Aristoteles Goes Neto, Coordenador(a) de curso de pós-graduação**, em 15/06/2022, às 10:47, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Fabrizio Murai Ferreira, Professor do Magistério Superior**, em 15/06/2022, às 17:55, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Jose Miguel Lima de Aguiar Graça Lourenço, Usuário Externo**, em 17/06/2022, às 12:26, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Luiz Carlos Junior Alcantara, Usuário Externo**, em 08/07/2022, às 16:55, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1495622** e o código CRC **34309AB9**.

Referência: Processo nº 23072.231724/2022-79

SEI nº 1495622

Agradecimentos

Agradeço à minha família, pelo suporte e amor em todos os momentos de dificuldade, por terem me ajudado a ver o caminho e me dado forças para trilhá-lo. À minha mãe, fonte inesgotável de sabedoria, fortaleza e amor. Meus irmãos Valério e Fernando, pela amizade companheirismo, me incentivam a ser sempre uma pessoa melhor. Ao meu pai, pelos momentos importantes e lições aprendidas ao acompanhá-lo no fim da sua jornada, acreditando mais em mim do que eu próprio e me ajudando a ver nosso processo nesta vida por uma perspectiva diferente. Aos meus avós pela bênção constante das suas presenças, que me preenchem de amor, ternura e paz.

Agradeço ao meu orientador, Prof. Dr. Luiz Carlos Junior Alcantara, por todo conhecimento e experiência compartilhados de maneira generosa. Sou grato também pelo voto de confiança e pela compreensão nos momentos de dificuldades pessoais.

Ao meu coorientador, Prof. Dr. José Lourenço, por acreditar no meu trabalho e contribuir de maneira excepcional ao meu crescimento científico.

Ao Prof. Dr. Vasco Ariston de Carvalho Azevedo, por me receber em seu laboratório de braços abertos e ter permitido o início dessa jornada.

À CAPES pelo financiamento da bolsa de estudos, fundamental para o alcance destes resultados.

Por fim, à equipe de pesquisadores da mais alta qualidade que compõe o projeto ZIBRA II, com a qual tive o prazer e a honra de trabalhar e aprender.

“O período de maior ganho em conhecimento e experiência
é o período mais difícil da vida de alguém.”

Dalai Lama

Resumo

Arbovírus transmitidos por mosquitos apresentam grandes desafios para a saúde pública, sendo responsáveis por epidemias, que causam significativos impactos no sistema de saúde dos países onde ocorrem, com um espectro diversificado de afecções clínicas. Dentre os arbovírus, os vírus da febre amarela e da dengue são de interesse especial em território brasileiro. O vírus da febre amarela é responsável pela doença mais severa transmitida por mosquitos nos trópicos, sendo que o Brasil enfrentou surtos recentes com altas taxas de mortalidade, em áreas onde o vírus não era reportado há décadas. Esses locais consistem em regiões urbanas densamente povoadas com grande prevalência de pessoas não vacinadas. O vírus da dengue, por sua vez, é uma ameaça que põe em risco um terço da população global, especialmente em locais onde seu principal vetor, mosquitos do gênero *Aedes*, é mais prevalente. No Brasil, especificamente, diferentes sorotipos vêm causando grandes surtos nas últimas décadas, de modo que a vigilância genômica desponta como importante medida de detecção precoce de vírus emergente e reemergentes, assim como para a investigação do seu comportamento dinâmico e disseminação. Nesse contexto, ferramentas de análise de dados baseadas em aprendizado de máquina mostram-se capazes de extrair informações úteis de grandes volumes de dados e auxiliar nos objetivos da vigilância genômica. Nos trabalhos aqui apresentados, analisamos sequências genômicas completas e quase completas de vírus da febre amarela e da dengue, associados a dados clínicos, laboratoriais, epidemiológicos, geográficos e temporais, no intuito de identificar assinaturas genéticas correlacionadas a características biológicas observadas. Como resultado, identificamos variações nucleotídicas não sinônimas associadas ao “*cycle threshold*” de amostras de febre amarela de primatas não humanos e à severidade de infecções por febre amarela em humanos, para as quais realizamos modelagem proteica estrutural *in-silico* e discutimos possíveis implicações biológicas. Além disso, identificamos também assinaturas genéticas que diferenciam duas linhagens de vírus da dengue do sorotipo 2 em um surto recente no Brasil, destacando a complementaridade e a concordância do método com a análise filogenética. Sendo assim, este trabalho apresenta uma abordagem inicial, versátil e rápida, para auxiliar a vigilância genômica em tempo real.

Palavras chave: Bioinformática, aprendizado de máquina, genômica, inteligência artificial, vírus da febre amarela, vírus da dengue.

Abstract

Arboviruses transmitted by mosquitoes present major challenges for public health and are responsible for epidemics which cause significant impacts on the health system of the countries where they occur, with a diverse spectrum of clinical conditions. Among them, yellow fever and dengue viruses are of special interest in the Brazilian territory. Yellow fever virus is responsible for the most severe disease transmitted by mosquitoes in the tropics, and Brazil has faced recent outbreaks with high mortality rates in areas where the virus had not been reported for decades. These sites consist of densely populated urban regions with a high prevalence of unvaccinated people. Dengue virus, in turn, is a threat that puts a third of the global population at risk, especially in places where its main vector, mosquitoes of the *Aedes* genus, is more prevalent. In Brazil, specifically, different serotypes have caused major outbreaks in recent decades, so that genomic surveillance comes up as an important measure for early detection of emerging and reemerging viruses, as well as for investigating their dynamic behavior and dissemination. In this context, data analysis tools based on machine learning are able to extract useful information from large volumes of data and assist in the objectives of genomic surveillance. In the work presented here, we analyzed complete and almost complete genomic sequences of yellow fever and dengue viruses, associated with clinical, laboratory, epidemiological, geographic and temporal data, in order to identify genetic signatures correlated with observed biological characteristics. As a result, we identified non-synonymous nucleotide variations associated with the cycle threshold of yellow fever samples from non-human primates and the severity of yellow fever infections in humans, for which we performed *in-silico* structural protein modeling and discussed possible biological implications. In addition, we also identified genetic signatures that differentiate strains of dengue virus serotype 2 in a recent outbreak in Brazil, highlighting the method's complementarity and accordance to phylogenetic analysis. Therefore, this work presents an initial, versatile and fast approach to assist in real-time genomic surveillance.

Keywords: Bioinformatics, machine learning, genomics, artificial intelligence, yellow fever virus, dengue virus.

Lista de Figuras

<i>Figura 1 – Ciclos de transmissão de arbovírus. (adaptado de FIGUEIREDO, 2019).....</i>	<i>14</i>
<i>Figura 2 – Genoma de Flavivirus (adaptado de PIERSON; DIAMOND, 2020).....</i>	<i>16</i>
<i>Figura 3 (adaptado de LUNDBERG; LEE, 2017) - Diagrama de contribuições SHAP.</i>	<i>26</i>

Lista de Abreviações

C - Proteína do nucleocapsídeo

CHIKV – Vírus da chikungunya

DENV – Vírus da dengue

E - Proteína do envelope

GWAS – Estudo de associação em genoma completo (“*genome wide association study*”)

M – Proteína da membrana

ML – Aprendizado de máquina (“*machine learning*”)

MLR – Regressão logística modificada

NHP - Primata não humano (“*non-human primate*”)

NS - Proteína não estrutural (“*Non structural*”)

ORF - Janela de leitura aberta (“*open reading frame*”)

pb – Pares de bases

PCR Ct – Limiar de ciclo de reação em cadeia de polimerase (“*cycle threshold*”)

prM - Proteína pré-membrana

RF – Floresta aleatória (“*random forest*”)

RNA - Ácido ribonucleico

SNV – Variação única de nucleotídeo (“*single nucleotide variants*”)

XGB – Amplificação de gradiente extrema (“*extreme gradient boosting*”)

YFV – Vírus da febre amarela

ZIKV – Vírus da Zika

Sumário

1	INTRODUÇÃO	14
1.1	ARBOVÍRUS	14
1.1.1	<i>Gênero Flavivirus</i>	<i>15</i>
1.2	APRENDIZADO DE MÁQUINA APLICADO À VIGILÂNCIA GENÔMICA	19
1.2.1	<i>Tipos de algoritmos de aprendizagem de máquina</i>	<i>21</i>
1.2.2	<i>Modelos utilizados na tese</i>	<i>23</i>
1.2.3	<i>Interpretação de Modelos - SHAP</i>	<i>25</i>
2	OBJETIVOS	28
2.1	OBJETIVO GERAL	28
2.2	OBJETIVOS ESPECÍFICOS	28
3	RESULTADOS.....	29
3.1	ARTIGO 1: “MACHINE LEARNING MODELS EXPLORING CHARACTERISTIC SINGLE-NUCLEOTIDE SIGNATURES IN YELLOW FEVER VIRUS”.....	29
3.2	ARTIGO 2: “MACHINE LEARNING AND IN-SILICO STRUCTURAL ANALYSIS AS A TOOL FOR THE RAPID IDENTIFICATION AND CHARACTERIZATION OF CIRCULATING DENV LINEAGES DURING BRAZILIAN EPIDEMICS”	49
4	DISCUSSÃO GERAL	65
5	CONCLUSÃO.....	69
	REFERÊNCIAS BIBLIOGRÁFICAS	71
	ANEXO	77
	ARTIGOS PUBLICADOS.....	77
	CAPÍTULO DE LIVRO	78

Prefácio

A vigilância genômica de patógenos emergentes e reemergentes tem se mostrado fundamental na detecção precoce de surtos, acompanhamento da sua dinâmica geográfica e temporal, identificação de variantes de interesse, planejamento de medidas preventivas e de controle e desenvolvimento de vacinas e medicamentos. Nesse contexto, arboviroses têm sido um dos principais agentes infecciosos a emergir ou reemergir, causando significativos impactos em saúde pública e econômicos nas populações afetadas. Como resposta a esse desafio, o Brasil vem se tornando referência mundial em vigilância genômica em tempo real de arboviroses (<https://www.zibraproject.org> e <https://www.zibra2project.org>), com importante contribuição para os bancos de dados globais e para as políticas públicas de saúde.

No entanto, a grande quantidade de dados gerada pelas iniciativas de vigilância genômica exigem métodos avançados de análise, de forma que técnicas de aprendizado de máquina apresentam-se como uma das abordagens mais promissoras nessa nova era de “*big data*”. Por isso, o presente trabalho se propõe a efetuar análises de bioinformática baseadas em aprendizado de máquina a dados de sequenciamento de vírus da febre amarela e da dengue, obtidos pelo grupo de pesquisa e disponíveis em bancos de dados de acesso livre, no intuito de extrair informações que possam contribuir com os esforços de vigilância genômica no Brasil.

A presente tese será constituída por seis subdivisões, conforme apresentadas abaixo. Na primeira seção, foi realizada uma breve revisão bibliográfica sobre os temas tratados, abordando aspectos relacionados aos vírus estudados e ao estado da arte dos métodos de aprendizado de máquina empregados. A segunda seção apresenta os objetivos gerais e específicos do trabalho. A terceira seção apresenta os principais resultados obtidos no desenvolvimento do doutorado, constituídos por dois artigos científicos, sendo um indexado no bioRxiv, em processo de submissão, e em processo de finalização.

- No primeiro artigo, é investigada a existência de assinaturas genéticas, em sequências de vírus da febre amarela, que possam estar relacionadas à variação da carga viral em infecções de primatas não humanos e à variação na mortalidade em infecções de humanos, em surtos recentes no Brasil. Efetua-se a modelagem estrutural computacional das proteínas nas quais foram identificadas variações não sinônimas de nucleotídeos, discutindo as possíveis implicações biológicas.

- No segundo artigo, são investigadas assinaturas genéticas que caracterizem e diferenciem diferentes linhagens de DENV2 (BR-4L1 e BR-4L2) cocirculantes durante o recente surto de dengue no Brasil, em 2019.

Na quarta seção é feita a discussão integrada dos trabalhos realizados, onde é destacada a relevância de ferramentas baseadas em aprendizado de máquina para a vigilância genômica em tempo real. Na quinta sessão apresenta-se na conclusão da tese, com exposição das metas propostas e alcançadas durante o período de doutorado. Finalmente, na sexta sessão, está presente o Apêndice, com listagem de trabalhos adicionais produzidos como colaborador durante o doutorado.

1 Introdução

1.1 Arbovírus

Nas últimas décadas, zoonoses têm sido responsáveis pela maioria das doenças infecciosas emergentes em humanos (JONES et al., 2008). Dentre elas, os vírus transmitidos por mosquitos se destacam como grandes desafios para a saúde pública. São considerados arbovírus aqueles vírus transmitidos por artrópodes (“*arthropod borne virus*”), constituindo um grupo de agentes infecciosos zoonóticos cujos principais vetores são mosquitos e carrapatos (WHO SCIENTIFIC GROUP ON ARBOVIRUSES, 1967). O conjunto de arbovírus engloba quatro famílias virais com características ecológicas semelhantes (*Bunyaviridae*, *Reoviridae*, *Togaviridae* e *Flaviviridae*), porém sem necessária relação taxonômica, com capacidade de causar doenças graves, como febres hemorrágicas, encefalites e meningites, figurando entre as principais causas de doenças infecciosas a emergir ou reemergir em todo o mundo (MUSSO; GUBLER, 2016; WU et al., 2019).

O ciclo de transmissão de arbovírus (Figura 1) envolve complexas relações entre os vetores artrópodes, reservatórios vertebrados, humanos e o meio ambiente. Em geral, os seres humanos são hospedeiros acidentais, com algumas exceções importantes como dengue (DENV), febre amarela (YFV) e chikungunya (CHIKV) (CHAMBERS et al., 1990; MUSSO; GUBLER, 2016).

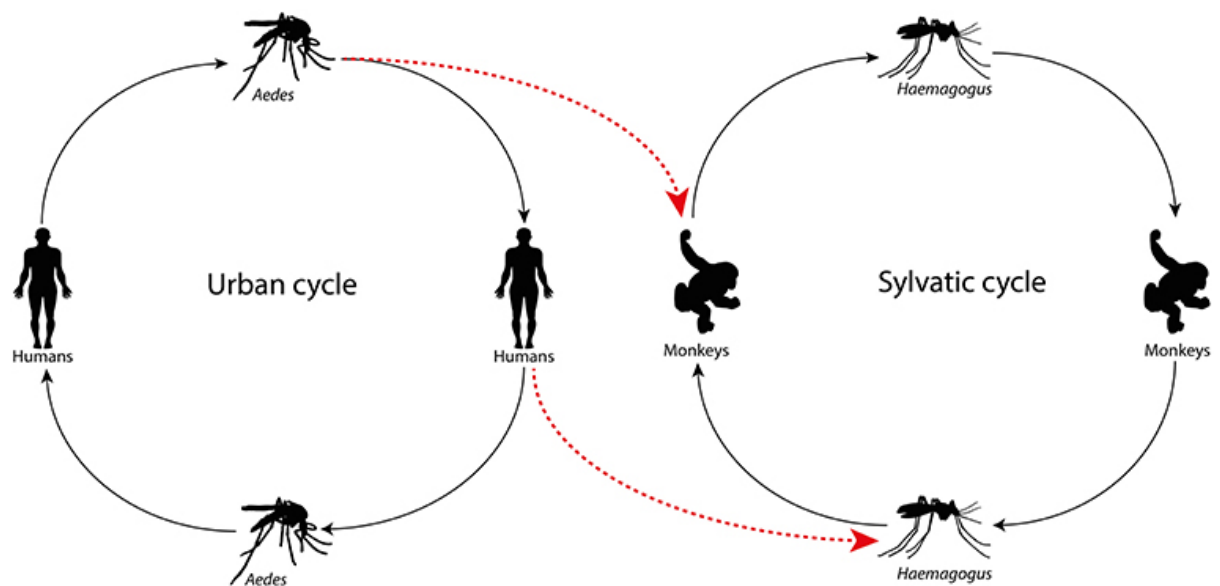


Figura 1 – Ciclos de transmissão de arbovírus. (adaptado de FIGUEIREDO, 2019)

Arboviroses tem sido responsáveis por epidemias que causam significativos impactos no sistema de saúde dos países onde ocorrem, com um espectro diversificado de afecções clínicas que vão desde indivíduos assintomáticos, quadros de doença febril de diferentes gravidades, erupções cutâneas, comprometimento articular, síndrome neurológica e síndrome hemorrágica (LOPES; NOZAWA; LINHARES, 2014). Além disso, o diagnóstico de arboviroses é complexo devido à semelhança de sinais e sintomas com outras doenças, à presença de casos assintomáticos ou com poucos sintomas, dificuldades na disponibilização de diagnóstico molecular ou sorológico diferencial e à possibilidade de reações cruzadas (MORELI; COSTA, 2013).

1.1.1 Gênero *Flavivirus*

A família *Flaviviridae* contém quatro gêneros, *Flavivirus*, *Pestivirus*, *Pegivirus* e *Hapacivirus*. Os arbovírus são encontrados somente dentro do gênero *Flavivirus*, que pode ser dividido em três grupos distintos: transmitidos por mosquitos, transmitidos por carrapatos e sem vetor definido (CLETON et al., 2012; HOLBROOK, 2017; SIMMONDS et al., 2017). Os vetores predominantemente associados à transmissão dos flavivírus são mosquitos do gênero *Aedes*, além de contribuição dos gêneros *Haemagogus* e *Sabethes* nas Américas (MUKTAR; TAMERAT; SHEWAFERA, 2016; VASCONCELOS; MONATH, 2016).

Flavivirus são vírus de RNA fita simples senso positivo, com aproximadamente 11Kb, que codificam uma única ORF (“*open reading frame*”), além de duas regiões não traduzidas 5’UTR e 3’UTR (“*untranslated region*”). O processo de tradução gera uma única poliproteína, posteriormente clivada em 10 proteínas (Figura 2), sendo três delas estruturais e sete não estruturais: C (capsídeo), M (membrana), E (envelope), NS1, NS2A, NS2B, NS3, NS4A, NS4B, NS5 (BARROWS et al., 2018; CHAMBERS et al., 1990; PIERSON; DIAMOND, 2020).

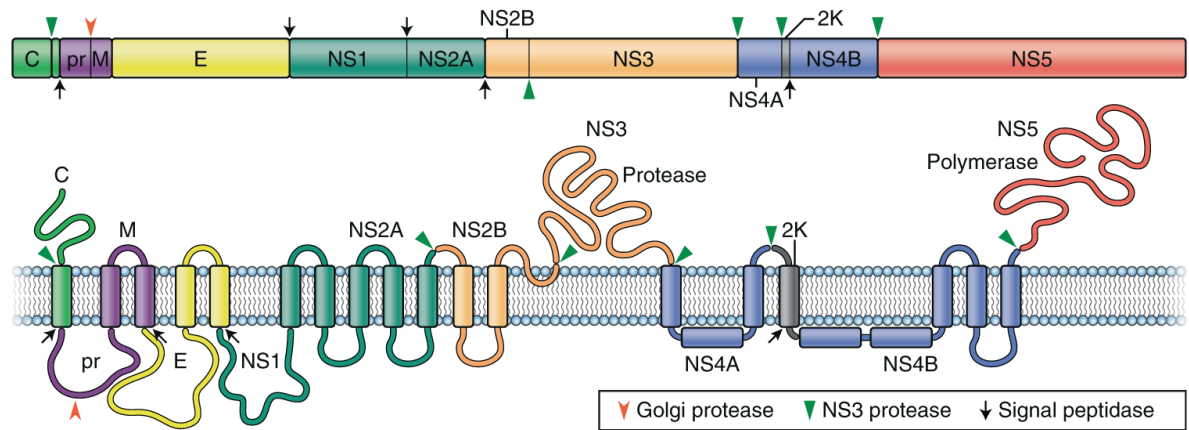


Figura 2 – Genoma de Flavivirus (adaptado de PIERSON; DIAMOND, 2020)

Diferentes flavivírus compartilham uma estrutura comum de partícula viral. A camada mais externa consiste em um envoltório glicoproteico, composto por múltiplas cópias da proteína E (envoltório) combinada à proteína M (membrana). Internamente, o genoma do vírus encontra-se complexado com cópias da proteína C (capsídeo), formando o nucleocapsídeo (BARROWS et al., 2018). As proteínas não estruturais participam da replicação do genoma viral, montagem das proteínas e estruturas virais, além de suprimir a resposta inata do hospedeiro à infecção (CHEN et al., 2017).

1.1.1.1 Febre Amarela

A febre amarela é uma doença hemorrágica viral aguda, endêmica em regiões tropicais da América do Sul e da África. Seu agente etiológico, o vírus da febre amarela (YFV), caracteriza o espécimen prototípico do gênero *Flavivirus*, com ORF composto por 10,233 pares de bases codificando uma única poliproteína (CHAMBERS et al., 1990; MONATH, 2001). A doença causada pelo YFV pode variar de um estado febril não específico até quadros de febre hemorrágica, muitas vezes fatais. Os sintomas geralmente surgem após um período de incubação de 3 a 6 dias seguindo a picada de um mosquito infectado, com o período de infecção durando vários dias. Em casos leves e abortivos de febre amarela, os pacientes podem se recuperar nesta fase sem sequelas duradouras. No entanto, após um período de remissão com duração de até 24h, aproximadamente 15-25% das pessoas infectadas entram em um período de intoxicação, quando a doença reaparece

de forma mais grave e as taxas de mortalidade atingem 20-50% (GARDNER; RYMAN, 2010; MONATH, 2001, 2008).

O tratamento da febre amarela é sintomático e a vacinação continua sendo o método de prevenção mais eficaz, fornecendo imunidade vitalícia em até 99% dos vacinados, de acordo com a Organização Mundial da Saúde (PAHO/WHO, [s.d.]). Embora extensas campanhas de vacinação combinadas com estratégias eficazes de controle de vetores tenham reduzido significativamente o número de casos de febre amarela em todo o mundo, estima-se que ocorram entre 84.000 a 170.000 casos graves e 29.000 a 60.000 mortes anualmente (PAULES; FAUCI, 2017; WHO, 2019), e cerca de 35 milhões de pessoas permanecem não vacinadas em áreas de risco no Brasil (SHEARER et al., 2017).

YFV pode se disseminar em dois ciclos diferentes: silvestre e urbano (Figura 1). O ciclo de transmissão silvestre ocorre em áreas florestais, onde o vírus é endemicamente transmitido entre várias espécies de primatas não humanos (NHP), que atuam como reservatório do vírus, e mosquitos vetores (notadamente *Haemagogus* spp. e *Sabethes* spp.) que habitam a região mais elevada da floresta, a copa das árvores. No ciclo silvestre, o homem é considerado hospedeiro acidental, infectando-se ao adentrar regiões de mata para atividades ocupacionais ou de lazer (MONATH, 2001). O ciclo de transmissão urbana ocorre quando o vírus é introduzido em áreas urbanizadas com alta densidade populacional, concomitantemente à presença acentuada de mosquitos urbanos, principalmente *Aedes aegypti*, onde o homem se torna o principal reservatório do vírus e a transmissão se estabelece de homem para homem através do mosquito vetor, resultando em epidemias explosivas e de grande escala (GARDNER; RYMAN, 2010). O ciclo urbano de transmissão do YFV foi erradicado no Brasil desde 1942, devido às campanhas de vacinação e controle de vetores (CONSOLI; DE OLIVEIRA, 1994; MARES-GUIA et al., 2020; VASCONCELOS, 2010).

Entretanto, durante a última década no Brasil, casos de febre amarela em humanos e epizootias acometendo primatas não humanos foram notificados em locais além dos limites da Amazônia brasileira, até então considerados endêmicos para a doença. Ocorreram duas grandes ondas de transmissão, uma durante o período sazonal 2016-2017, com 778 casos humanos confirmados, incluindo 262 mortes, e outra durante o período sazonal 2017-2018, com 1.376 casos

humanos confirmados, incluindo 483 mortes. Como resultado, desde 2020, o Brasil mudou sua recomendação de vacinação contra febre amarela para incluir todo o país (PAHO/WHO, 2021b).

Existem sete genótipos diferentes de YFV identificados até o presente momento, sendo cinco pertencentes à África (África Ocidental I e II, África Oriental, África Oriental/Central e Angola) e duas presentes nas Américas (América do Sul I e II), ambas derivadas do genótipo da África Ocidental (BRYANT; HOLMES; BARRETT, 2007). O genótipo predominante no Brasil é o América do Sul I, com 5 linhagens diferentes (1A-1E), sendo que as recentes epidemias de YFV no Brasil foram causadas pelas linhagens mais recentes (D e E), que substituíram as mais antigas (SILVA et al., 2020).

O diagnóstico da febre amarela é realizado através de métodos virológicos (nomeadamente, detecção do genoma viral, detecção de antígenos virais ou isolamento do vírus) e/ou sorológicos (ELISA, PRNT) (PAHO/WHO, 2018).

1.1.1.2 Dengue

O vírus da dengue (DENV) é uma ameaça global, representando desafios de saúde para um terço da população mundial, especialmente em locais onde seu principal vetor, mosquitos do gênero *Aedes*, são mais prevalentes (KRAEMER et al., 2015). O DENV reemergiu nas últimas décadas principalmente devido a uma maior distribuição geográfica dos mosquitos vetores, juntamente com o aumento da mobilidade humana e crescimento populacional, urbanização não planejada, globalização e mudanças climáticas. Além disso, a falta de métodos eficazes de controle de vetores, vacinas ou tratamentos antivirais contribuem para a reemergência do DENV. Hoje, o DENV circula em regiões tropicais e subtropicais em todo o mundo, com um impacto que cresceu de 1,5 milhão de casos acumulados na década de 1980 para 16,2 milhões na década de 2010-2019 (BARROWS et al., 2018; PAHO/WHO, 2021a; SAN MARTÍN et al., 2010).

A infecção por dengue é normalmente subclínica. Porém, pode apresentar-se como uma doença autolimitada que cursa com febre alta, fortes dores no corpo e outros sintomas gripais. A progressão para dengue severa é rara, mas pode resultar em doença hemorrágica grave, com consequentes afecções respiratórias, sangramento, choque e óbito (BARROWS et al., 2018).

Como outros integrantes do gênero *Flavivirus*, DENV possui um genoma viral de RNA de fita simples senso positivo com aproximadamente 11kb. Seu genoma codifica uma única poliproteína que é posteriormente clivada nas 10 proteínas típicas do gênero.

O DENV é classificado em quatro sorotipos virais intimamente relacionados, mas antigenicamente distintos (DENV1, DENV2, DENV3 e DENV4), cada um abrangendo diferentes genótipos. A vigilância da diversidade viral circulante é de extrema importância para a saúde pública, pois os casos mais graves de dengue ocorrem após segunda infecção por um sorotipo heterotípico (MARTINEZ et al., 2020). Além disso, o impacto na resposta imune, causado por infecções decorrentes de diferentes genótipos do mesmo sorotipo, ainda não foi totalmente elucidado, o que reitera a relevância da vigilância genômica e de estudos nesse sentido.

Diferentes sorotipos causaram grandes epidemias no Brasil nos últimos 20 anos, com a reemergência do DENV2 no ano de 2019 causando uma epidemia grave, com 1.544.987 casos prováveis e 782 óbitos confirmados (até a semana epidemiológica – EW52 2019) (MINISTÉRIO DA SAÚDE DO BRASIL, 2019, 2020).

Embora o Brasil ainda não tenha atingido sucesso durador em seus esforços de erradicação do vetor de DENV, em especial mosquitos *Aedes aegypti*, o país vem se tornando referência mundial em vigilância genômica em tempo real de arboviroses (<https://www.zibraproject.org> e <https://www.zibra2project.org>), com importante contribuição para os bancos de dados globais, assim como para o melhor conhecimento a respeito da dinâmica viral, evolutiva e espaço-temporal de arbovírus.

1.2 Aprendizado de Máquina aplicado à Vigilância Genômica

A vigilância genômica é fundamental para a detecção precoce de vírus emergentes e reemergentes, assim como para o entendimento do comportamento dinâmico desses agentes e sua disseminação. Nesse sentido, o sequenciamento de genomas virais em tempo real desempenha um papel importante na luta contra as epidemias emergentes e reemergentes. Os genomas gerados e analisados fornecem informações que permitem caracterizar a diversidade genética viral, reconstruir as origens da epidemia, estimar as taxas de transmissão e prever próximos surtos. Além disso, o conhecimento dos genomas virais circulantes é fundamental para o desenvolvimento de

vacinas e novos medicamentos, além de contribuir para o aperfeiçoamento dos métodos diagnósticos sorológicos e moleculares (FARIA et al., 2016, 2017).

Para tanto, é necessária a combinação entre infraestrutura com capacidade de geração de grande volume de dados genômicos e recursos adequados para a análise desse grande volume de dados. Porém, a análise de grandes conjuntos de dados genômicos e sua correlação com as características observadas não é tarefa fácil, exigindo abordagens estatísticas e computacionais robustas e sofisticadas.

Nesse contexto, conjuntos de dados genômicos de grande escala apresentam desafios significativos para as análises estatísticas e de bioinformática típicas, pois a alta dimensionalidade das características genômicas, bem como o número geralmente baixo de amostras disponíveis, torna as abordagens clássicas não apropriadas para o problema. Além disso, a estrutura altamente correlacionada dos dados genômicos viola a suposição de independência exigida pela maioria dos modelos estatísticos padrão. Em problemas de análise de dados como esses, geralmente são aplicadas abordagens de aprendizado de máquina (ML) (CHEN; ISHWARAN, 2012; ISHWARAN et al., 2010).

Conforme descrito em trabalhos que analisam a eficácia de estudos de associação em genoma completo (“*genome-wide association studies*” - GWAS) em larga escala, as relações não lineares entre variantes genéticas e características observadas, bem como o impacto geralmente baixo de variações individuais nos resultados observados, exigem poderosos métodos de mineração de dados e aprendizado de máquina. Essas abordagens são utilizadas para modelar computacionalmente a relação entre combinações de variantes de nucleotídeo único (“*single nucleotide variants*” - SNVs), outras variações genéticas e fatores ambientais com resultados observados. Além disso, a seleção de quais SNVs considerar em uma análise, levando em conta suas relações não lineares, requer algum tipo de filtragem, redução de dimensionalidade ou quantificação da importâncias de SNVs individuais (BEHRAVAN et al., 2018; HO et al., 2019; MOORE; ASSELBERGS; WILLIAMS, 2010).

Ademais, a interpretação biológica desses modelos de ML é de extrema importância para aplicações práticas (MOORE; ASSELBERGS; WILLIAMS, 2010). A capacidade de interpretar corretamente a saída de um modelo é de grande interesse, pois cria confiança no usuário, fornece informações sobre como um modelo pode ser aprimorado e, de particular importância em

problemas de Bioinformática, auxilia na compreensão do processo subjacente que está sendo modelado (BAO; ZHOU; CUI, 2005; BEHRAVAN et al., 2018; HO et al., 2019; LUNDBERG et al., 2018; LUNDBERG; LEE, 2017; RIBEIRO; SINGH; GUESTRIN, 2016). Quando um modelo de ML pode ser interpretado, é possível avaliar, por exemplo, quais variantes genéticas estão relacionadas com os fenótipos em estudo, possibilitando conhecer os fundamentos biológicos dos padrões observados nos dados. É por isso que, em muitas aplicações, modelos simples (por exemplo, modelos lineares) são frequentemente preferidos por sua facilidade de interpretação. No entanto, a crescente disponibilidade de dados aumentou os benefícios do uso de modelos complexos, que são mais opacos à interpretação. Assim, deve-se obter um equilíbrio entre complexidade e interpretabilidade.

Levando esses desafios em consideração, algoritmos baseados em árvore e “*boosting*” destacam-se como especialmente passíveis de interpretação, embora ainda robustos e capazes de lidar com problemas de alta dimensão, variáveis interdependentes e tamanho amostral pequeno (QI, 2012).

1.2.1 Tipos de algoritmos de aprendizagem de máquina

Tradicionalmente, a utilização de computadores para realização de tarefas envolve a construção de um conjunto de regras inequívocas e definidas, que codificam as etapas sequenciais para que se solucione o problema em questão. Algoritmos de aprendizado de máquina, por sua vez, utilizam-se de dados para identificar padrões e fazer inferências. Nesse processo, pode-se dizer que o algoritmo é treinado e tem seus parâmetros internos ajustados de acordo com os dados observados.

Entretanto, como é natural, existem diversas abordagens diferentes para se implementar algoritmos de aprendizagem de máquina, cada uma mais adequada a tipos diferentes de dados e a tarefas ou perguntas diferentes que se busca responder. Podem ser divididos, inicialmente, entre supervisionados, não supervisionados, semi supervisionados e de reforço. Algoritmos supervisionados recebem dados de treinamento para os quais a classificação é conhecida, ao passo que os não supervisionados recebem dados cuja classificação é desconhecida e tentam identificar padrões, por exemplo, por meio de agrupamento de semelhantes. Semi supervisionados utilizam-

se de uma mistura de dados conhecidos e não conhecidos. Algoritmos por reforço buscam diversas soluções para um problema na tentativa de maximizar uma função de recompensa.

No presente trabalho, como foram utilizadas somente sequências genômicas que possuíam classificação conhecida, como será detalhado adiante, os algoritmos utilizados enquadram-se no tipo supervisionado.

Além disso, algoritmos de aprendizado de máquina são agrupados em cinco grandes grupos, de acordo com (DOMINGOS, 2015): simbolistas, conexionistas, evolucionistas, Bayesianos e analogistas.

Algoritmos simbolistas utilizam-se do raciocínio indutivo/dedutivo, construindo o conhecimento a partir de características observadas nos dados. Um exemplo de algoritmos simbolista é árvore de decisão, incluindo seus refinamentos, como *random forest* e *boosting*.

Algoritmos conexionistas baseiam-se na estrutura neurológica, composta por neurônios conectados por sinapses, através dos quais a informação é processada. Exemplos incluem redes neurais artificiais, que utilizam *backpropagation* para ajustar os pesos das sinapses de acordo com os dados de treinamento. São particularmente opacas à interpretação, em especial as que possuem muitas camadas, como *deep learning*.

A abordagem evolucionista se baseia no processo de evolução por seleção natural. A partir de uma função de *fitness* a ser otimizada, o algoritmo realiza iterações sucessivas de uma população de possíveis soluções, aplicando mutações e cruzamento, até que as melhores soluções sejam alcançadas. São muito úteis em problemas de otimização e na construção de algoritmos que evoluem e tornam-se mais adequados às tarefas com o tempo.

Algoritmos baseados na abordagem Bayesiana utilizam ferramentas estatísticas para inferir soluções que tenham a maior probabilidade de estarem corretas, ao passo em que quantificam a chance de que alternativas possíveis também sejam a resposta para o problema em questão.

Por fim, algoritmos analogistas utilizam-se de padrões internos nos dados e os comparam a padrões conhecidos, criando analogias e classificando os dados em grupos conhecidos. Um exemplo de implementação seria máquina de vetor de suporte.

Uma observação final sobre a aplicação de modelos de ML é que, devido à falta de distinção a priori entre algoritmos de aprendizagem (WOLPERT, 1996), é sempre aconselhável comparar

diferentes soluções de ML ao abordar um problema pela primeira vez. Portanto, no trabalho incluído nesta tese apresento a aplicação de três algoritmos de ML, descritos a seguir, agregando seus resultados para interpretação.

1.2.2 Modelos utilizados na tese

1.2.2.1 *Random Forest*

Entre os modelos baseados em árvores, o “*Random Forest*” (RF) (BREIMAN, 2001) é um dos mais utilizados e possui aplicações muito amplas em problemas de mineração de dados e aprendizado de máquina. Seu sucesso no campo da bioinformática deve-se ao fato de que a RF pode ser aplicada a uma ampla gama de tipos de dados, mesmo quando o problema que está sendo modelado não é linear ou envolve efeitos complexos de alta ordem, como classificação de expressão gênica, espectrometria de massa, proteômica ou análise metabolômica, descoberta de biomarcadores, anotação de sequência, previsão de interação proteína-proteína ou genética estatística, tornando-o particularmente atraente para análise de dados genômicos (QI, 2012). É um método altamente adaptável aos dados, pode ser aplicado a problemas de alta dimensionalidade com poucas amostras e é capaz de explicar correlações e interações entre variáveis, o que é especialmente útil para dados genômicos de alta dimensão (BAO; ZHOU; CUI, 2005; CHEN et al., 2007; CHEN; ISHWARAN, 2012; MOORE; ASSELBERGS; WILLIAMS, 2010; QI, 2012; WANG et al., 2005).

O algoritmo consiste em um agregado de árvores de decisão, ou preditores, de tal forma que cada árvore é construída a partir de um conjunto aleatório de amostras de entrada, escolhidas independentemente e com a mesma distribuição para todas as árvores (também conhecido como “*bootstrapping and aggregating - bagging*”). Além disso, cada nó de cada árvore é calculado com base em um subconjunto aleatório das características das amostras, eliminando assim a correlação entre as árvores, bem como possibilitando a generalização do modelo, à medida que mais árvores são adicionadas. A criação das árvores de decisão de maneira sequencial, com a divisão em nós seguindo uma lógica “*greedy*”, permite que a regularização seja imposta naturalmente, possibilitando análise efetiva em problemas com grande dimensionalidade e limitado número de amostras (BREIMAN, 2001). Além disso, a “propriedade de agrupamento” de árvores

(ISHWARAN et al., 2010) permite que RF lide habilmente com correlação e interação entre variáveis.

Outra vantagem da RF é que, durante o treinamento, algumas amostras são deixadas de fora do processo de “*bootstrapping*”. Essas amostras “*out-of-bag*” (OOB) podem ser usadas como um conjunto de validação para medir o desempenho do modelo à medida que ele é treinado, evitando assim a necessidade de um conjunto de dados de validação separado. Essa medida de desempenho pode ser usada para comparar diferentes modelos de RF no mesmo conjunto de dados e decidir quantas árvores usar para cada problema de classificação.

1.2.2.2 *Extreme Gradient Boosting*

De acordo com (SCHAPIRE, 2003), o método “*boosting*” parte do princípio de que soluções para problemas de regressão e classificação podem ser baseadas em um conjunto de regras fracas (preditores fracos), no lugar de uma única regra altamente poderosa. Na verdade, determinar um número grande o suficiente de pequenas regras práticas e rudimentares, combinando-as, pode ser muito mais fácil do que encontrar uma única regra de previsão altamente precisa. Em outras palavras, as técnicas de “*boosting*” executam um método geral e comprovadamente eficaz de produzir uma regra de previsão muito precisa, combinando preditores moderadamente imprecisos de maneira sequencial, cada vez alimentando o modelo com um conjunto de exemplos de treinamento com pesos individuais diferentes. Para cada iteração, o algoritmo de aprendizado básico gera uma nova regra de previsão fraca, combinando essas regras fracas em uma única regra de previsão (conjunto preditor forte) após muitas iterações.

Nesse contexto, “*gradient boosting*” (FRIEDMAN, 2001) é uma aplicação de “*boosting*” que se utiliza do gradiente dos resíduos do modelo como função objetivo a ser minimizada, em cada etapa de treinamento. Ele é um método de aprendizado de máquina amplamente utilizado, também baseado em árvores de decisão, que apresenta ótimo desempenho em diversas aplicações, fornecendo excelentes resultados em uma ampla gama de referências padronizadas de classificação (LI, 2010). Conforme descrito por (LUBKE et al., 2013) em problemas de estudo de variação genômica, o método não requer a especificação a priori de um modelo genético, além de permitir a inclusão de um grande número de variáveis não lineares e não independentes. Atualmente, a implementação mais utilizada é o XGBoost (“*Extreme Gradient Boosting*”), por ser eficiente e

escalável, disponível como um pacote de código aberto para Python (<https://github.com/dmlc/xgboost>) (CHEN; GUESTRIN, 2016).

1.2.2.3 Regressão Logística Regularizada

Finalmente, os modelos clássicos de regressão logística são um método padrão usado em bioinformática para classificação de dados e previsão de doenças. Uma de suas principais vantagens são os resultados facilmente interpretáveis, onde os valores dos coeficientes estão diretamente relacionados às importâncias das variáveis correspondentes. No entanto, os modelos clássicos de regressão logística não convergem para uma solução única quando o conjunto de dados contém mais variáveis do que amostras, como é o caso em análise baseado em informação genética. Uma das abordagens mais utilizadas para superar esse problema é aplicar técnicas de redução de dimensionalidade, como Análise de Componentes Principais (PCA), reduzindo o número de variáveis a um pequeno conjunto não correlacionadas linearmente, com alta variância. Outra solução é aplicar a regularização ao modelo, de modo que os parâmetros sejam forçados a terem valores restritos e pequenos, e aqueles com pouco efeito no resultado do modelo fiquem naturalmente próximos de zero. Dessa forma, têm sido propostas técnicas de regularização, que introduzem um termo de penalidade à função a ser minimizada, no sentido de ajustar analiticamente um modelo de regressão logística ao conjunto de dados de alta dimensão, denominado “regressão logística modificada” (MLR) (MORAIS-RODRIGUES, 2017).

1.2.3 Interpretação de Modelos - SHAP

Técnicas de aprendizado de máquina usam métodos estatísticos para inferir relações entre atributos de dados (variáveis independentes) e resultados (variáveis dependentes) em grandes conjuntos de dados. No presente trabalho, são buscadas relações entre variantes genéticas e fenômenos biológicos resultantes. No entanto, a complexidade da maioria dos problemas do mundo real, em particular os biológicos, torna as previsões de ML difíceis de interpretar e, portanto, com limitada possibilidade de ação.

Métodos interpretáveis explicam por que uma determinada previsão foi feita para um caso particular, ou seja, as características genéticas específicas que levaram à previsão, tornando

possível obter informações sobre os fundamentos biológicos e os padrões observados nos dados. É por isso que, em muitas aplicações, modelos simples (por exemplo, modelos lineares) são frequentemente preferidos por sua facilidade de interpretação. No entanto, a crescente disponibilidade de “*big data*” aumentou os benefícios do uso de modelos complexos, que são mais robustos e precisos, porém mais opacos à interpretação. Assim, deve-se obter um equilíbrio entre precisão e interpretabilidade.

Com essas dificuldades em mente, vários métodos têm sido propostos como formas de interpretar modelos de ML, em especial para modelos baseados em árvores de decisão (BACH et al., 2015; DATTA; SEN; ZICK, 2016; LIPOVETSKY; CONKLIN, 2001; RIBEIRO; SINGH; GUESTRIN, 2016; SHRIKUMAR; GREENSIDE; KUNDAJE, 2017; ŠTRUMBELJ; KONONENKO, 2014). No entanto, muitas vezes não está claro como esses métodos estão relacionados entre si e quando um método é preferível aos outros. Para lidar com esse problema, foi proposto um framework unificado para interpretação de previsões chamado SHAP (SHapley Additive exPlanations) (LUNDBERG et al., 2018; LUNDBERG; LEE, 2017, 2018), que é utilizado neste trabalho.

SHAP é uma abordagem unificada para interpretar previsões de modelos com base em Teoria dos Jogos (valores "Shapley") que calcula contribuições aditivas de cada variável independente no modelo final, e garante uma solução única, cujos valores SHAP são grandezas escalares que representam uma medida de importância de cada variável no resultado dos modelos considerados. Quando SHAP é aplicado a um modelo e seu conjunto de dados, o método calcula a importância relativa de cada variável na contribuição para a classificação ou regressão realizada pelo modelo, para cada amostra, explicando o impacto que cada uma delas tem na previsão, para essa amostra em particular. Isso pode ser melhor visualizado na Figura 3 abaixo:

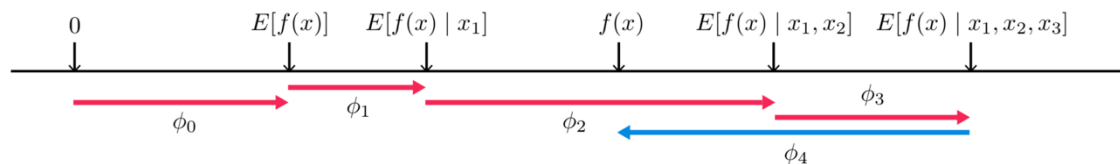


Figura 3 (LUNDBERG; LEE, 2017) - Diagrama de contribuições SHAP – Os valores SHAP (ϕ_i) atribuem a cada variável independente (x_i) uma contribuição no valor esperado da predição do modelo para aquela amostra, quando condicionada àquela variável. Assim, eles explicam como parte-se do valor esperado base $E[f(x)]$ que seria predito caso não se conhecesse nenhuma das variáveis da amostra, para a predição final $f(x)$. Dessa forma, a saída do modelo é explicada como a soma das contribuições de cada variável.

No contexto da interpretabilidade do modelo, métodos de “comitê” ou de conjunto de preditores, como XGBoost e RF, são um caso especial para o qual foi desenvolvida uma implementação dedicada, rápida e exata, para calcular os valores SHAP, chamada Tree SHAP (LUNDBERG; LEE, 2018). Essa implementação existe sobre um pacote Python, que possibilita aplicá-la diretamente a RF e ao XGBoost. O código está disponível gratuitamente online (<http://github.com/slundberg/shap>).

2 Objetivos

2.1 Objetivo geral

Desenvolver uma estrutura analítica baseada em aprendizado de máquina, passível de ser aplicada a qualquer tipo de genoma viral, na busca por assinaturas genéticas relacionadas a características de interesse, como variações em Ct de PCR (limiar de ciclo de PCR) de amostras de YFV de primatas não humanos, severidade de infecção por YFV em casos humanos, variações características de nucleotídeos e aminoácidos em linhagens cocirculantes de DENV2, identificadas por análise filogenética. Esperando com isto contribuir com as ações de vigilância genômica em tempo real no Brasil.

2.2 Objetivos específicos

- Definir características biológicas de interesse a serem investigadas;
- Realizar o pré-tratamento de dados genômicos, clínicos e epidemiológicos para deixá-los adequados à análise por algoritmos de aprendizado de máquina, gerando novos pipelines computacionais para a tarefa;
- Buscar por assinaturas genéticas em sequências de YFV que estejam relacionadas à variação de Ct em sequências provenientes de primatas não humanos;
- Buscar por assinaturas genéticas em sequências de YFV que estejam relacionadas à severidade da doença em sequências provenientes de humanos;
- Buscar por assinaturas genéticas em sequências de DENV2 que sejam características e diferenciem cada uma das linhagens cocirculantes BR-4L1 e BR-4L2;
- Realizar modelagem e análise estrutural de proteínas de YFV e DENV, identificando onde ocorrem as variações não sinônimas identificadas, discutindo possíveis implicações biológicas de acordo com simulações *in-silico* e literatura atual;
- Demonstrar utilidade da aplicação da metodologia desenvolvida conjuntamente às demais análises rotineiramente realizadas em vigilância genômica em tempo real, como filogenia;

3 Resultados

3.1 Artigo 1: “Machine Learning models exploring characteristic single-nucleotide signatures in Yellow Fever Virus”

O vírus da febre amarela (YFV) é o agente da mais grave doença transmitida por mosquitos nos trópicos dada a sua taxa de mortalidade alta. Recentemente, o Brasil sofreu grandes surtos de YFV, afetando regiões onde o vírus não era relatado há décadas, consistindo em áreas urbanas onde vive um grande número de pessoas não vacinadas. Nesta tese foi desenvolvida uma estrutura de aprendizado de máquina combinando três algoritmos diferentes (XGBoost, Random Forest e Regressão Logística Regularizada). Este método foi aplicado a 56 sequências de YFV provenientes de infecções humanas e 27 de infecções de primatas não humanos (NHPs), para investigar a presença de assinaturas genéticas possivelmente relacionadas à gravidade da doença (em sequências relacionadas a humanos) e diferenças nos valores do limiar do ciclo de PCR (Ct) (em sequências relacionadas com NHP).

As análises revelam quatro variações nucleotídicas (SNVs) não sinônimas em sequências de infecções humanas, nas proteínas NS3 (E614D), NS4a (I69V), NS5 (R727G, V643A) e seis SNVs não sinônimas em sequências NHP, nas proteínas E (L385F), NS1 (A171V), NS3 (I184V) e NS5 (N11S, I374V, E641D). Foi ainda realizada uma análise estrutural comparativa de proteínas nesses SNVs, descrevendo possíveis impactos na função da proteína.

Apesar do conjunto de dados ser limitado em tamanho e de este estudo não considerar as interações vírus-hospedeiro, o trabalho desenvolvido destaca o uso do aprendizado de máquina como uma abordagem inicial versátil e rápida para a exploração de dados genômicos.

Machine learning models exploring characteristic single-nucleotide signatures in Yellow Fever Virus

Álvaro Salgado^{a,†*}, Raquel C. de Melo-Minardi^{b*}, Marta Giovanetti^{a,c*}, Adriano Veloso^{b*}, Francielly Moraes-Rodrigues^a, Talita Adelino^d, Ronaldo de Jesus^e, Stephane Tosta^a, Vasco Azevedo^a, Jose Lourenço^{f#}, Luiz Carlos J. Alcantara^{a,c#}

^a Laboratório de Genética Celular e Molecular, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

^b Departamento de Ciência da Computação, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

^c Laboratório de Flavivírus, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil

^d Laboratório Central de Saúde Pública, Fundação Ezequiel Dias, Belo Horizonte, Minas Gerais, Brazil

^e Coordenação Geral dos Laboratórios de Saúde Pública, Secretaria de Vigilância em Saúde, Ministério da Saúde, Brasília, DF, Brazil

^f Department of Zoology, University of Oxford, Oxford OX1 3PS, UK

Running Head: Machine learning applied to YFV genomic investigation

*These authors contributed equally.

#Corresponding authors: Álvaro Salgado, alvarosalgado@ufmg.br; Luiz Carlos J Alcantara, luiz.alcantara@ioc.fiocruz.br; Jose Lourenço, jose.lourenco@zoo.ox.ac.uk

Abstract

Yellow fever virus (YFV) is the agent of the most severe mosquito-borne disease in the tropics. Recently, Brazil suffered major YFV outbreaks with a high fatality rate affecting areas where the virus has not been reported for decades, consisting of urban areas where a large number of unvaccinated people live. We developed a machine learning framework combining three different algorithms (XGBoost, random forest and regularized logistic regression). This method was applied to 56 YFV sequences from human infections and 27 from non-human primate (NHPs) infections to investigate the presence of genetic signatures possibly related to disease severity (in human related sequences) and differences in the PCR cycle threshold (Ct) values (in NHP related sequences). Our analyses reveal four non-synonymous single nucleotide variations (SNVs) on sequences from human infections, in proteins NS3 (E614D), NS4a (I69V), NS5 (R727G, V643A) and six non-synonymous SNVs on NHP sequences, in proteins E (L385F), NS1 (A171V), NS3 (I184V) and NS5 (N11S, I374V, E641D). We performed comparative protein structural analysis on these SNVs, describing possible impacts on protein function. Despite the fact that the dataset is limited in size and that this study does not consider virus-host interactions, our work highlights the use of machine learning as a versatile and fast initial approach to genomic data exploration.

Importance

Yellow fever is responsible for 29-60 thousand deaths annually in South America and Africa and is the most severe mosquito-borne disease in the tropics. Given the range of clinical outcomes and the availability of YFV genomic data, the use of machine learning analysis promises to be a powerful tool in the investigation of genetic signatures that could impact disease severity and its potential of being reintroduced in an urban transmission cycle. This can assist in the search for biomarkers of severity as well as help elucidating variations in host's Ct value. This work aims to propose a relatively fast and inexpensive computational analysis framework, which can be used as a real-time, initial strategy associated with genomic surveillance to identify a set of single nucleotide variants putatively related to biological and clinical characteristics being observed.

Introduction

Yellow fever (YF) is an acute viral hemorrhagic disease endemic in tropical areas of Africa and Latin America. The causative agent, yellow fever virus (YFV), represents the prototypical member of the genus *Flavivirus* (family *Flaviviridae*), consisting of a single-stranded, positive-sense RNA virus, with a genome about 11,000 kb and a single open-reading frame of 10,233 nucleotides (1, 2). Disease varies from nonspecific febrile illness to a fatal hemorrhagic fever. Symptoms usually appear after an incubation period of three to six days following the bite of an infected mosquito, with a period of infection lasting several days (2–4). The World Health Organization (WHO) reports case fatality rates in the order of 15 to 50% (5). Vaccination remains the most effective YF prevention method, providing lifetime immunity in up to 99% of vaccinated people (6). Nevertheless, the burden of YF is estimated to be between 84,000 to 170,000 severe cases and 29,000 to 60,000 deaths annually (7, 8), while an estimated 35 million people remain unvaccinated in areas at risk in Brazil only (9).

YFV spreads in two different cycles: sylvatic and urban. The sylvatic transmission cycle occurs in forested areas, where the virus is endemically transmitted between several non-human primate (NHP) species. The urban transmission cycle occurs when the virus is introduced into human populations with high density and urban-dwelling mosquitoes (mainly *Aedes aegypti*) (3). Urban cycles of YFV transmission have been eradicated in Brazil since 1942 due to vaccination and vector control campaigns (10–13).

In the last decade in Brazil, however, human and NHP epizootic YF cases have been notified at places beyond the limits of regions previously considered (sylvatic) endemic for the virus (14–17). The severe impact of these recent outbreaks can be measured, in part, by its fatality rate at around 34%, higher than the general rate estimated by Monath and colleagues (4), motivating the inquiry as to what could be the possible factors contributing to such a high fatality rate, and if YFV genetic signatures could be among those factors.

Additionally, important findings in recent epidemics (11, 18) show a significant difference in the distribution of NHP Ct values, in which *Callithrix* spp. exhibit generally higher Ct values than other NHP species, do not develop fatal YFV infections similar to those reported in humans and can persist for longer, thus increasing the infectious period. The latter can be an essential factor in

igniting an urban cycle of transmission, mainly due to the genus' proximity to densely urbanized areas.

On this respect, genomic and epidemiological monitoring have become an integral part of the national (Brazil) and international response to emerging and ongoing epidemics of viral infectious diseases, allowing the availability of a large amount of genomic data (19–23).

In genomic and epidemiological monitoring analysis, machine learning (ML) approaches are usually applied (24, 25), as described in works that analyze the effectiveness of large scale genome-wide association studies (GWAS), due to their capability to computationally model the relationship between combinations of single nucleotide variants, other genetic variations and environmental factors with observed outcomes (26–28).

We curated two different datasets of YFV genomes, one from human cases and the other from NHP cases. After data curation, the human dataset contained 56 YFV sequences, with 40 sequences related to infections leading to severe outcomes or death, and 16 sequences related to cases with no severe outcome. We also gathered an NHP (*Callithrix* spp.) dataset, that after curation contained 27 sequences, of which 21 were related to low Ct values (< 20) and 6 were related to high Ct values (≥ 20).

We applied three different ML models to each dataset, to guarantee robustness of the ML analysis (29). We then analysed the models using SHAP (SHapley Additive explanation) (30–32) to highlight genetic signatures. The possible biological impacts of these signatures were investigated and discussed by means of in-silico protein structural analysis coupled with literature review.

Results

Non-human primates Ct value statistical analysis

Figure 1 shows the distribution of cycle threshold (Ct) values from *Callithrix* spp. sequences, with two distinct clusters roughly around 12 and 30, with a median value of 26.1. The result of Hartigan's dip test of unimodality (33) rejected the null hypothesis of a unimodal distribution for Ct values ($p < .001$), which indicate the existence of two groups of *Callithrix* spp. Ct values.

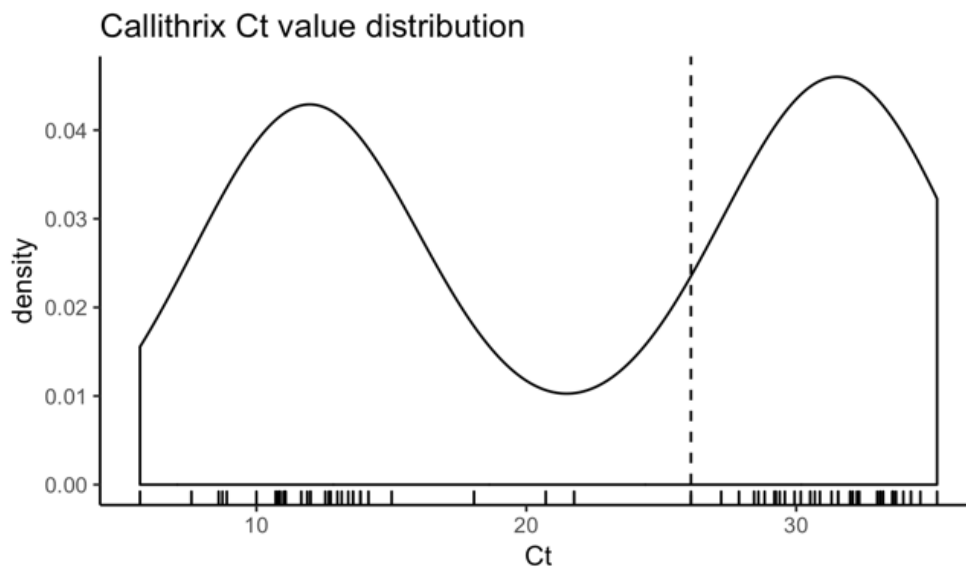


Figure 1 – Ct values associated with *Callithrix* spp. sequences. The median value shown by a dashed line. Hartigan's dip test of unimodality indicates bimodal distribution ($p < .001$).

Machine learning models' performance

Figure 2 shows the confusion matrices for the machine learning models applied. For the human dataset, the XGBoost classification model achieved an accuracy of 75% on the test set, with F-1 scores of 0.59 and 0.82 for classes 0 (not severe/not death) and 1 (severe/death), respectively. The random forest model achieved an accuracy of 82% on the test set, with F-1 scores of 0.74 and 0.86 for classes 0 and 1 respectively. The modified logistic regression model achieved an accuracy of 82% on the test set, with F-1 scores of 0.74 and 0.86 for classes 0 and 1, respectively.

For the *Callithrix* spp. dataset, the three models achieved an accuracy of 100% on the test set, with F-1 scores of 1.00 and 1.00 for classes 0 (low Ct) and 1 (high Ct) respectively.

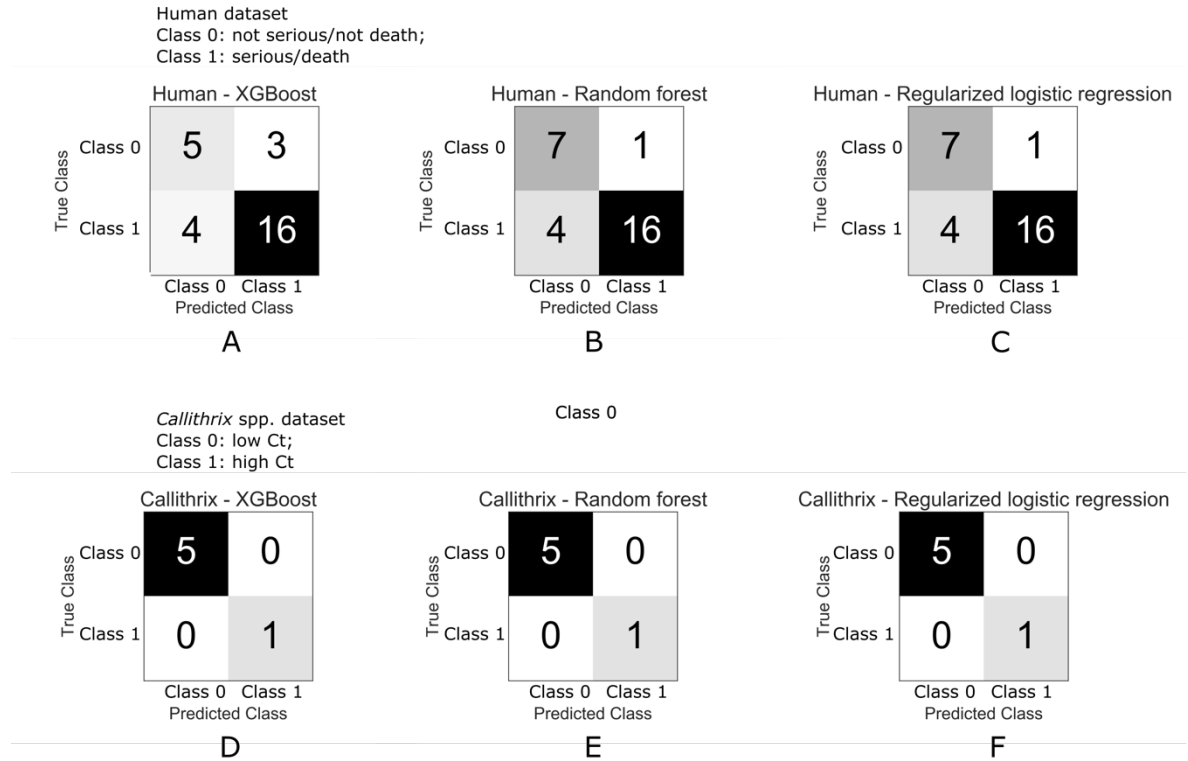


Figure 2 - Confusion matrices. Top three figures correspond to human test dataset and the bottom three figures correspond to *Callithrix* spp. test dataset, for XGBoost, random forest and regularized logistic regression respectively.

YFV genetic signatures

The machine learning methods identified the non-synonymous single nucleotide variants (SNVs) shown in Table 1. The table displays each protein where the SNV was found, with nucleotide position on YFV genome, position relative to the protein's sequence, amino acid position relative to the translated protein, reference genome amino acid and corresponding codon, analyzed sequences amino acid variation and corresponding codon and SNV position inside codon (1st, 2nd or 3rd).

The results obtained by the analysis of human YFV sequences highlighted 4 SNV positions that result in the amino acid change, and the results obtained by the analysis of *Callithrix* spp. shows 6 SNV positions that resulted in the amino acid change.

For each of these SNVs, variability between sequences was measured via Shannon entropy.

Table 1 - YFV genetic signatures

Human dataset									
Protein	nn position	nn position on protein	aa position on protein	aa reference	codon reference	aa variation	codon variation	SNV codon position (1, 2, 3)	Entropy
NS3	6412	1842	614	E	gaa	D D	gac gat	3	0.34/0.00
NS4a	6644	205	69	I	atc	V	gtc	1	0.70/0.47
NS5	9815	2179	727	R	agg	G	ggg	1	0.00/0.17
NS5	9564	1928	643	V	gtt	A	gct	2	0.70/0.00
NHP dataset									
Protein	nn position	nn position on protein	aa position on protein	aa reference	codon reference	aa variation	codon variation	SNV codon position (1, 2, 3)	Entropy
NS5	8756	1120	374	I	atc	V	gtc	1	0.45/0.00
NS5	9559	1923	641	E	gaa	D D	gac gat	3	0.28/0.00
NS3	5120	550	184	I	atc	V	gtc	1	0.28/0.00
E	2126	1153	385	L	ctc	F	ttc	1	0.28/0.00
NS5	7647	11	4	N	aat	S	agt	2	0.28/0.00
NS1	2964	512	171	A	gca	V	gta	2	0.28/0.00

Protein structural analysis

We performed protein structural analysis for all SNVs indicated in Table 1. The templates, their resolution, the quality of models provided from Swiss-Model (34) and the changes in binding affinity and stability predicted by mCSM-NA (35) are summarized in Table 2

Table 2 - Protein structural analysis results

	Callithrix spp. Dataset						Human dataset			
Protein	E	NS1	NS3	NS5			NS3	NS4a	NS5	
SNV	L385F	A171V	I184V	N11S	I374V	E641D	E614D	I69V	R727G	V643A
Template (PDB ID)	Template (6WI5) (?)	Zika virus NS1 (5K6K) (27)	Dengue virus NS3 (5YV8:A) (31)	Yellow fever virus NS5 (6QSN) (32)	Yellow fever virus NS5 (6QSN) (32)	Yellow fever virus NS5 (6QSN) (32)	Yellow fever virus NS3 (1YKS) (29)	No structure deposited in the PDB	Yellow fever virus NS5 (6QSN) (32)	Yellow fever virus NS5 (6QSN) (32)
Resolution	1.83 Å	1.89 Å	2.5 Å	3.00 Å	3.00 Å	3.00 Å	1.80 Å	-	3.00 Å	3.00 Å
Coverage (%)		100%	99%					-		
Sequence identity		47.58%	50.57%					-		
Localization at protein	Domain III	Wing flexible loop	Linker region	MTase domain	Palm subdomain	Palm subdomain	Helicase domain	-	Thumb subdomain	Palm subdomain
Global Model Quality Estimation (GMQE)		0.79	0.76					-		
(Quality mean) QMEAN		-2.5	-2.22					-		
Predicted change in binding affinity	Target distant from binding sites				$\Delta\Delta G = 0.003$ Kcal/mol (Increased affinity)		$\Delta\Delta G = 0.025$ Kcal/mol (Increased affinity)	-	$\Delta\Delta G = -1.545$ Kcal/mol (Reduced affinity)	$\Delta\Delta G = -0.001$ Kcal/mol (Reduced affinity)
Predicted change in stability					-1.908 Kcal/mol (Destabilising)		-0.254 Kcal/mol (Destabilising)	-	-0.751 Kcal/mol (Destabilising)	-1.884 Kcal/mol (Destabilising)

Figure 3 shows the structural representation of proteins E, NS1, NS3 and NS5, with corresponding SNVs.

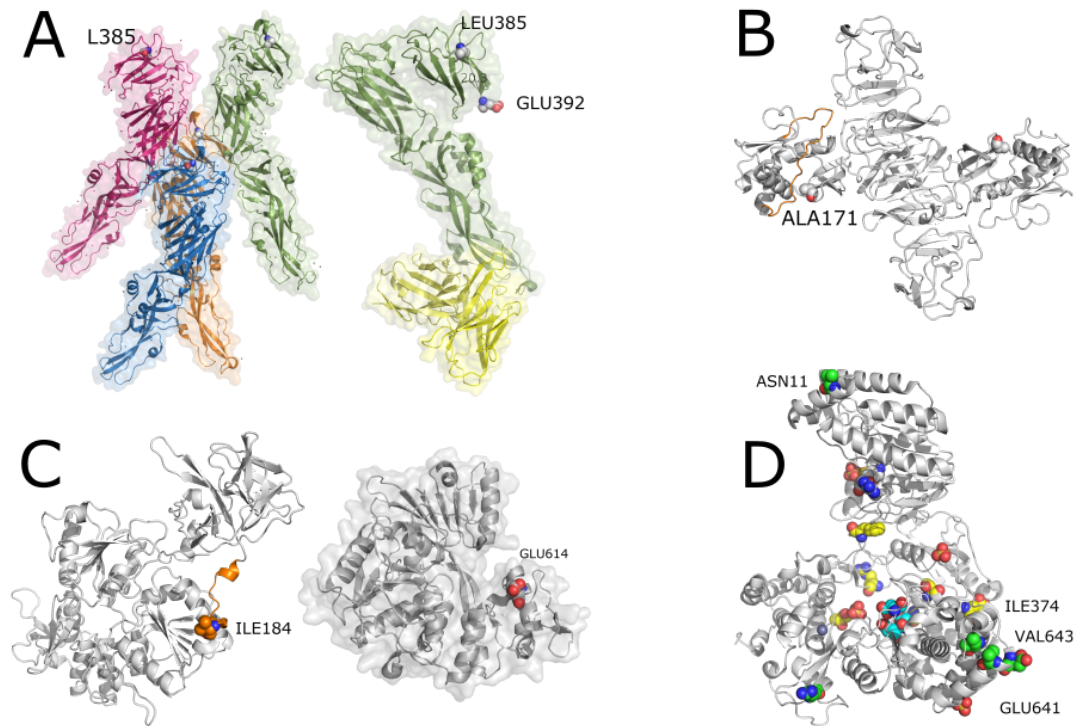


Figure 3 - (A) Envelope protein - (A, left) PDB id 6wi5 - tetramer of E protein LEU385 presented in spheres. (A, right) PDB id 6ivz - in green, one chain of protein E and in yellow, the light and heavy chains of monoclonal antibody 5A. Note that the L385F SNV is distant from both the binding between protein E chains and the antibody recognition site. (B) NS1 protein. Comparative model built with Swiss-Model and PDB id 5k6k. In orange, we depict the wing flexible loop and in spheres, SNV ALA171. (C) NS3 protein. (C, left) Comparative model built with Swiss-Model and PDB id 5yvu. In orange, we depict the interdomain linker region and in spheres, SNV ILE184. (C, right) GLU614 from PDB id 1yks showing that the SNV E614D is close to the cleft where the DNA binds the NS3 protein. (D) NS5 protein. In green, SNVs. In yellow, important conserved residues across ZIKV, DENV and WNV. In cyan, active site. In grey, ligand S-adenosyl-L-homocysteine. Sulfate and Zn ions are also represented in spheres.

E (envelope) protein

LEU385 (NHP dataset) is far from the intra-chain binding site and the antibody recognition site (Figure 3-A). LEU385 is in the Domain III (DIII) which has an immunoglobulin C domain (IgC-like) presenting a seven-stranded fold and is supposed to contain the receptor-binding site. DIII suffers a rotation and goes closer to the fusion loop (FL), bringing the C-terminal part of DIII (residue 392) close to FL. LEU385 is 20.3 angstroms far from GLU392.

NS1 protein

The SNV A171V (NHP dataset) is in close contact with a region called *wing flexible loop* (highlighted in orange in Figure 3-B) and it is also a probable glycosylation site (36).

NS3 protein

Although there is a crystallographic structure (PDB id: 1yks) of the helicase domain (37) and another containing part of NS3 complexed with NS2B (PDB id 6urv) (38) deposited in the PDB, the referred SNV occurs in the unresolved stretch. The I184V (NHP dataset) is in the linker region (shown in orange in Figure 3-C, left). It connects protease and helicase domains and corresponds to sequence KEEGKEELQEIP that encompasses residues between 174 and 185. The SNV E614D (human dataset) occurs in the helicase domain and is located in the RNA binding cleft, shown in red on Figure 3-C, right.

NS4a protein

Since it has multiple transmembrane hydrophobic segments, structural analysis of NS4a has been unsuccessful and, so far, there is no structure deposited in the PDB. It is still one of the least characterized proteins from YFV. It was not possible to obtain a good structural model since the best template found had a coverage of only 37% and a sequence identity of 25.53%.

NS5 protein

There is a recent YFV NS5 structure deposited on PDB (PDB id 6qsn) (39). The analyzed SNVs (shown in green in Figure 3-D) are N11S, the only SNV in the MTase domain; I374V and E641D, both located in the palm subdomain. These three SNVs were found on the NHP dataset. Additionally, SNV V643A, located in the palm, and R727G, in the thumb subdomain, were found in the human dataset. As depicted in Figure 3-D, they (green) are located far from the Zn and sulfate ions and the ligand (S-adenosyl-L-homocysteine) (grey). They are not close to any important / conserved mentioned residue (yellow) that interact with the nucleic acid. SNV I374V is also present in ZIKV, DENV and WNV. Position E641D varies across other viruses (K, N, R). Position V643A is also not conserved being an insertion, K or N. Position R727G is S, E or T in other flaviviruses.

Discussion

In this study, we demonstrate the potential of applying ML approaches on real-time genomic surveillance, to quickly identify genetic loci which may be of public health interest.

We find signals in multiple genetic loci and present a structural-based review on the potential impact of changes at those loci. However, the limited number of sequences analyzed demands caution when presenting the results. A large number of high-quality sequences is ideal for the application of ML analysis, especially when dealing with viruses, whose high mutation rates tend to insert many variations on its genomes. Furthermore, our analysis didn't consider virus-host interactions, such as host genome or immune system and pre-existing health conditions. In this regard, efficient host data collection, such as Electronic Health Records (EHR), are of paramount importance for a thorough investigation of clinical outcomes.

The envelope (E) protein is related to virus attachment and fusion (40). NHP dataset analysis shows SNV L395F on Domain III, a region containing an IgC-like domain and supposed to contain a receptor-binding site crucial for virion maturation (41). It is possible that this SNV could have an impact on the plasticity of this domain and affect the virus' receptor-binding site and it would be interesting to investigate this behaviour through simulations of molecular dynamics in future work.

NS1 protein is a crucial non-structural protein (36). We found a non-synonymous SNV (A171V) on NHP dataset, located near a highly flexible region on the protein, called the *wing flexible loop*, which is a probable glycosylation site (GS) (42). NS1 is also a key protein secreted by infected cells, which has the potential to interact with the adaptive immune system responses (36). In dengue infections, NS1 is known to modulate capillary leakage in severe disease and may thus have a role to play in the severity of YFV infection (43).

NS3 protein, which is composed of protease and helicase domains, has functions related to viral polyprotein processing and cleavage, viral genome replication and RNA capping (40). SNV I184V, found on NHP dataset, is in a region with probable limited functional constraints (44). SNV E614D, found on the human dataset, occurs in the NS3 helicase domain and is located in the RNA binding cleft. ASP and GLU are both negatively charged amino acids, but ASP has a shorter side chain which can cause it to lose access to the ligand. We used mCSM-NA (35) to evaluate the impact of the SNV on stability and affinity with RNA, showing a small destabilizing

effect on the interaction with RNA (-0.646 Kcal/mol), which could have an impact on its function, fundamental to viral genome replication. Unfortunately, there are no structures in complex with RNA available. Models built with protein-RNA docking techniques could help elucidate if there is a significant impact of this SNV on RNA interaction.

We found one SNV on the NS4a protein (I69V, human dataset). However, a lack of current knowledge on YFV NS4a impeded us from further exploring the possible role of I69V in human hosts. Based on the protein's proposed functions (45–47), this SNV could, in principle, affect viral replication, but such hypothesis would have to be tested by non-computational means.

As an outlier, NS5 had the highest number of identified variations - I374V, E641D, N11S, R727G, V643A - from both human and NHP YFV sequences. NS5 protein is a fundamental enzyme for viral replication because it contains an N-terminal methyltransferase domain (MTase) and a C-terminal RNA dependent RNA polymerase domain (RdRp) (48). MTase domain has important functions involved in protecting viral RNA from degradation and innate immunity response. RdRp is essential for viral RNA replication, because its activities cannot be performed by host polymerases, and is a promising target for antiviral drug development (49). Furthermore, dengue virus NS5 has been associated with immune response evasion (50). With the structural analyses, we found that none of the SNVs found has been reported to be in positions with apparent connections to protein function or structure. However, R727G on the human dataset, on the thumb subdomain of RdRp domain, shows a change in predicted affinity for RNA upon SNV occurrence. This reduction in affinity could impact polymerase function and viral replication efficiency.

In conclusion, even though the method proposed was applied on data that was already available from other sources, our study demonstrate that it is efficient and easy to replicate, making it suitable for real-time genomic surveillance, in which genetic data is analyzed as it is generated. This approach may help detect and inform on possible connections between ongoing genetic changes and public health in a timely manner.

Materials and Methods

Datasets

We retrieved YFV complete or near complete genome sequences from recent Brazilian outbreaks, available on public databases (19–23), with associated epidemiological and clinical data containing relevant information regarding clinical severity and outcome for human infections samples, as well as PCR cycle threshold value (Ct) for both human and NHP samples. The multifasta file was aligned to a reference sequence (NC_002031.1) using MAFFT online (51) and was manually verified and corrected using AliView (<https://ormbunkar.se/aliview/>). This ensures the correct column numbering and, later, allows interpretation of results. After alignment, the reference sequence was removed.

Dataset rows contain the genome sequences analyzed, with each column containing the corresponding nucleotide on that genomic position. Each genomic position is, therefore, a data feature, and it has four possible values (A, C, G and T).

Sequences with coverage lower than 90% were removed from the study. Afterwards, any column that contained an empty position in any of the sequences was entirely removed. Finally, columns that did not contain nucleotide variation (i.e. uninformative) were removed. The initial NHP dataset contained 36 *Callithrix* sequences with 10080 genomic positions, and the initial human dataset contained 92 sequences with 10354 genomic positions. After data cleaning described above, NHP dataset was left with 27 sequences and 86 genomic positions. Human dataset was left with 56 sequences and 151 genomic positions.

Since the features are categorical (four possible nucleotides), “one-hot encoding” was applied to binarize the data.

For ML analysis, human infection sequences were divided between not severe/not death and severe/death. *Callithrix* spp. infection sequences were divided by low Ct (<20) and high Ct ($\geq$ 20). After curation, out of the 56 human YFV sequences, 40 were related to severe/death cases, and 16 sequences related to not severe/not death cases. The NHP (*Callithrix* spp.) dataset, after curation, 21 sequences related to low Ct values (< 20) and 6 related to high Ct values ($\geq$ 20).

Machine learning model adjustment

Three ML models were applied for each of two analyzed datasets, XGBoost (52, 53), random forest (54) and regularized logistic regression (55). XGBoost model parameters were adjusted in a “grid-search cross-validation” scheme with five folds. Random forest adjustment used “out of bag” data as validation. Regularized logistic regression parameters were adjusted on a 10-fold cross-validation scheme, using their averages in the final model.

Model interpretation

Feature importance was computed using SHAP (SHapley Additive exPlanation) (30–32). The author’s suggestion was followed (<https://github.com/slundberg/shap/issues/397>) on dealing with categorical data when using SHAP.

Protein structural analysis

We searched on Protein Data Bank (PDB) (56) for experimentally resolved structures. For those proteins that did not have structures, we looked for templates for comparative modeling with at least 30% identity. The comparative models were built with the Swiss-Model server (57). We used mCSM (35) method to predict the impact of SNVs on protein stability and interactions.

Acknowledgements

This work was supported by Decit, SCTIE, Brazilian Ministry of Health, Conselho Nacional de Desenvolvimento Científico - CNPq - (440685/2016-8 and 440856/2016-7), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES (88887.130716/2016-00, 88881.130825/2016-00 and 88887.130823/2016-00), by the European Union’s Horizon 2020 Research and Innovation Programme under ZIKAlliance Grant Agreement No. 734548.

References

1. Chambers TJ, Hahn CS, Galler R, Rice CM. 1990. Flavivirus genome organization, expression, and replication. *Annu Rev Microbiol* 44:649–688.
2. Monath TP. 2001. Yellow fever: an update. *Lancet Infect Dis* 1:11–20.
3. Gardner CL, Ryman KD. 2010. Yellow Fever: A Reemerging Threat. *Clinics in Laboratory Medicine* 30:237–260.
4. Monath TP. 2008. Treatment of yellow fever. *Antiviral Research* 78:116–124.
5. 2015. WHO Report on Global Surveillance of Epidemic-prone Infectious Diseases - Yellow fever. World Health Organization. World Health Organization. https://www.who.int/csr/resources/publications/yellowfev/CSR_ISR_2000_1/en/. Retrieved 2 July 2020.
6. 2020. Yellow Fever. Pan American Health Organization / World Health Organization. https://www.paho.org/hq/index.php?option=com_topics&view=article&id=69&Itemid=40784&lang=en. Retrieved 29 June 2020.
7. Paules CI, Fauci AS. 2017. Yellow Fever — Once Again on the Radar Screen in the Americas. *New England Journal of Medicine* 376:1397–1399.
8. 2019. Yellow fever. World Health Organization. <https://www.who.int/news-room/fact-sheets/detail/yellow-fever>. Retrieved 1 July 2020.
9. Shearer FM, Moyes CL, Pigott DM, Brady OJ, Marinho F, Deshpande A, Longbottom J, Browne AJ, Kraemer MU, O'Reilly KM, others. 2017. Global yellow fever vaccination coverage from 1970 to 2016: an adjusted retrospective analysis. *The Lancet infectious diseases* 17:1209–1217.
10. Consoli RA, de Oliveira RL. 1994. Principais mosquitos de importância sanitária no Brasil. SciELO-Editora FIOCRUZ.
11. Mares-Guia MAM de M, Horta MA, Romano A, Rodrigues CDS, Mendonça MCL, dos Santos CC, Torres MC, Araujo ESM, Fabri A, de Souza ER, Ribeiro ROR, Lucena FP, Junior LCA, da Cunha RV, Nogueira RMR, Sequeira PC, de Filippis AMB. 2020. Yellow fever epizootics in non-human primates, Southeast and Northeast Brazil (2017 and 2018). *Parasites & Vectors* 13.
12. 2017. Yellow fever, the return of an old threat. Fiocruz. <https://portal.fiocruz.br/en/news/yellow-fever-return-old-threat>. Retrieved 1 July 2020.
13. Vasconcelos PF da C. 2010. Yellow fever in Brazil: thoughts and hypotheses on the emergence in previously free areas. *Revista de Saúde Pública* 44:1144–1149.
14. Delatorre E, Abreu FVS de, Ribeiro IP, Gómez MM, dos Santos AAC, Ferreira-de-Brito A, Neves MSAS, Bonelly I, de Miranda RM, Furtado ND, Raphael LMS, da Silva L de FF, de Castro

MG, Ramos DG, Romano APM, Kallás EG, Vicente ACP, Bello G, Lourenço-de-Oliveira R, Bonaldo MC. 2019. Distinct YFV Lineages Co-circulated in the Central-Western and Southeastern Brazilian Regions From 2015 to 2018. *Front Microbiol* 10.

15. DIVE - Boletim Epidemiológico nº 06/2020 Situação epidemiológica da Febre Amarela em Santa Catarina (Atualizado em 10/06/2020). <http://www.dive.sc.gov.br/index.php/arquivo-noticias/1204-boletim-epidemiologico-n-06-2020-situacao-epidemiologica-da-febre-amarela-em-santa-catarina-atualizado-em-10-06-2020>. Retrieved 6 January 2021.

16. Boletim epidemiológico da Febre Amarela no Brasil 2019/2020 | RETS - Rede Internacional de Educação de Técnicos em Saúde. <http://www.rets.epsjv.fiocruz.br/biblioteca/boletim-epidemiologico-da-febre-amarela-no-brasil-20192020>. Retrieved 6 January 2021.

17. 2020. Febre Amarela | Secretaria da Saúde. <https://www.saude.pr.gov.br/Editoria/Febre-Amarela>. Retrieved 6 January 2021.

18. Cunha MS, da Costa AC, de Azevedo Fernandes NCC, Guerra JM, dos Santos FCP, Nogueira JS, D'Agostino LG, Komninakis SV, Witkin SS, Ressio RA, Maeda AY, Vasami FGS, Kaigawa UMA, de Azevedo LS, de Souza Facioli PA, Macedo FLL, Sabino EC, Leal É, de Souza RP. 2019. Epizootics due to Yellow Fever Virus in São Paulo State, Brazil: viral dissemination to new areas (2016–2017). *Scientific Reports* 9:5474.

19. Cunha M dos P, Duarte-Neto AN, Pour SZ, Ortiz-Baez AS, Černý J, Pereira BB de S, Braconi CT, Ho Y-L, Perondi B, Sztajn bok J, Alves VAF, Dolhnikoff M, Holmes EC, Saldiva PHN, Zanotto PM de A. 2019. Origin of the São Paulo Yellow Fever epidemic of 2017–2018 revealed through molecular epidemiological analysis of fatal cases. *Scientific Reports* 9.

20. Faria NR, Kraemer MUG, Hill SC, Jesus JG de, Aguiar RS, Iani FCM, Xavier J, Quick J, Plessis L du, Dellicour S, Thézé J, Carvalho RDO, Baele G, Wu C-H, Silveira PP, Arruda MB, Pereira MA, Pereira GC, Lourenço J, Obolski U, Abade L, Vasylyeva TI, Giovanetti M, Yi D, Weiss DJ, Wint GRW, Shearer FM, Funk S, Nikolay B, Fonseca V, Adelino TER, Oliveira M a. A, Silva MVF, Sacchetto L, Figueiredo PO, Rezende IM, Mello EM, Said RFC, Santos DA, Ferraz ML, Brito MG, Santana LF, Menezes MT, Brindeiro RM, Tanuri A, Santos FCP dos, Cunha MS, Nogueira JS, Rocco IM, Costa AC da, Komninakis SCV, Azevedo V, Chieppe AO, Araujo ESM, Mendonça MCL, Santos CC dos, Santos CD dos, Mares-Guia AM, Nogueira RMR, Sequeira PC, Abreu RG, Garcia MHO, Abreu AL, Okumoto O, Kroon EG, Albuquerque CFC de, Lewandowski K, Pullan ST, Carroll M, Oliveira T de, Sabino EC, Souza RP, Suchard MA, Lemey P, Trindade GS, Drumond BP, Filippis AMB, Loman NJ, Cauchemez S, Alcantara LCJ, Pybus OG. 2018. Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science* 361:894–899.

21. Giovanetti M, de Mendonça MCL, Fonseca V, Mares-Guia MA, Fabri A, Xavier J, de Jesus JG, Gräf T, dos Santos Rodrigues CD, dos Santos CC, Sampaio SA, Chalhoub FLL, de Bruycker Nogueira F, Theze J, Romano APM, Ramos DG, de Abreu AL, Oliveira WK, do Carmo Said RF, de Alburque CFC, de Oliveira T, Fernandes CA, Aguiar SF, Chieppe A, Sequeira PC, Faria NR, Cunha RV, Alcantara LCJ, de Filippis AMB. 2019. Yellow Fever Virus Reemergence and Spread in Southeast Brazil, 2016–2019. *Journal of Virology* 94.

22. Goes de Jesus J, Gräf T, Giovanetti M, Mares-Guia MA, Xavier J, Lima Maia M, Fonseca V, Fabri A, dos Santos RF, Mota Pereira F, Ferraz Oliveira Santos L, Reboredo de Oliveira da Silva L, Pereira Gusmão Maia Z, Gomes Cerqueira JX, Thèze J, Abade L, Cordeiro M de CS, Torquato SSC, Santana EB, de Jesus Silva NS, Dourado RSO, Alves AB, do Socorro Guedes A, da Silva Filho PM, Rodrigues Faria N, de Albuquerque CFC, de Abreu AL, Martins Romano AP, Croda J, do Carmo Said RF, Cunha GM, da Fonseca Cerqueira JM, Mello AL e S de, de Filippis AMB, Alcantara LCJ. 2020. Yellow fever transmission in non-human primates, Bahia, Northeastern Brazil. *PLOS Neglected Tropical Diseases* 14:e0008405.
23. Hill SC, de Souza R, Thézé J, Claro I, Aguiar RS, Abade L, Santos FCP, Cunha MS, Nogueira JS, Salles FCS, Rocco IM, Maeda AY, Vasami FGS, du Plessis L, Silveira PP, de Jesus JG, Quick J, Fernandes NCCA, Guerra JM, Réssio RA, Giovanetti M, Alcantara LCJ, Cirqueira CS, Díaz-Delgado J, Macedo FLL, Timenetsky M do CST, de Paula R, Spinola R, Telles de Deus J, Mucci LF, Tubaki RM, de Menezes RMT, Ramos PL, de Abreu AL, Cruz LN, Loman N, Dellicour S, Pybus OG, Sabino EC, Faria NR. 2020. Genomic Surveillance of Yellow Fever Virus Epizootic in São Paulo, Brazil, 2016 – 2018. *PLOS Pathogens* 16:e1008699.
24. Chen X, Ishwaran H. 2012. Random forests for genomic data analysis. *Genomics* 99:323–329.
25. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. 2010. High-Dimensional Variable Selection for Survival Data. *Journal of the American Statistical Association* 105:205–217.
26. Behravan H, Hartikainen JM, Tengström M, Pylkäs K, Winqvist R, Kosma V, Mannermaa A. 2018. Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls. *Scientific Reports* 8.
27. Ho DSW, Schierding W, Wake M, Saffery R, O’Sullivan J. 2019. Machine Learning SNP Based Prediction for Precision Medicine. *Frontiers in Genetics* 10.
28. Moore JH, Asselbergs FW, Williams SM. 2010. Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 26:445–455.
29. Wolpert DH. 1996. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation* 8:1341–1390.
30. Lundberg S, Lee S-I. 2017. A Unified Approach to Interpreting Model Predictions. *arXiv:170507874 [cs, stat]*.
31. Lundberg SM, Lee S-I. 2018. Consistent feature attribution for tree ensembles. *arXiv:170606060 [cs, stat]*.
32. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DK-W, Newman S-F, Kim J, Lee S-I. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* 2:749–760.
33. Hartigan JA, Hartigan PM. 1985. The Dip Test of Unimodality. *Ann Statist* 13:70–84.

34. Arnold K, Bordoli L, Kopp J, Schwede T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 22:195–201.
35. Pires DEV, Ascher DB. 2017. mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. *Nucleic Acids Res* 45:W241–W246.
36. Brown WC, Akey DL, Konwerski J, Tarrasch JT, Skinotis G, Kuhn RJ, Smith JL. 2016. Extended Surface for Membrane Association in Zika Virus NS1 Structure. *Nat Struct Mol Biol* 23:865–867.
37. Wu J, Bera AK, Kuhn RJ, Smith JL. 2005. Structure of the Flavivirus Helicase: Implications for Catalytic Activity, Protein Interactions, and Proteolytic Processing. *J Virol* 79:10268–10277.
38. Noske GD, Gawriljuk VO, Fernandes RS, Furtado ND, Bonaldo MC, Oliva G, Godoy AS. 2020. Structural characterization and polymorphism analysis of the NS2B-NS3 protease from the 2017 Brazilian circulating strain of Yellow Fever virus. *Biochim Biophys Acta Gen Subj* 1864:129521.
39. Dubankova A, Boura E. 2019. Structure of the yellow fever NS5 protein reveals conserved drug targets shared among flaviviruses. *Antiviral Research* 169:104536.
40. Barrows NJ, Campos RK, Liao K-C, Prasanth KR, Soto-Acosta R, Yeh S-C, Schott-Lerner G, Pompon J, Sessions OM, Bradrick SS, Garcia-Blanco MA. 2018. Biochemistry and Molecular Biology of Flaviviruses. *Chemical Reviews* 118:4448–4482.
41. Lu X, Xiao H, Li S, Pang X, Song J, Liu S, Cheng H, Li Y, Wang X, Huang C, Guo T, ter Meulen J, Daffis S, Yan J, Dai L, Rao Z, Klenk H-D, Qi J, Shi Y, Gao GF. 2019. Double Lock of a Human Neutralizing and Protective Monoclonal Antibody Targeting the Yellow Fever Virus Envelope. *Cell Reports* 26:438-446.e5.
42. Watanabe Y, Bowden TA, Wilson IA, Crispin M. 2019. Exploitation of glycosylation in enveloped virus pathobiology. *Biochimica et Biophysica Acta (BBA) - General Subjects* 1863:1480–1497.
43. Puerta-Guardo H, Glasner DR, Harris E. 2016. Dengue Virus NS1 Disrupts the Endothelial Glycocalyx, Leading to Hyperpermeability. *PLOS Pathogens* 12:e1005738.
44. Luo D, Xu T, Hunke C, Grüber G, Vasudevan SG, Lescar J. 2008. Crystal Structure of the NS3 Protease-Helicase from Dengue Virus. *Journal of Virology* 82:173–183.
45. Zou J, Xie X, Wang Q-Y, Dong H, Lee MY, Kang C, Yuan Z, Shi P-Y. 2015. Characterization of Dengue Virus NS4A and NS4B Protein Interaction. *Journal of Virology* 89:3455–3470.
46. Miller S, Kastner S, Krijnse-Locker J, Bühler S, Bartenschlager R. 2007. The non-structural protein 4A of dengue virus is an integral membrane protein inducing membrane alterations in a 2K-regulated manner. *J Biol Chem* 282:8873–8882.

47. Lin M-H, Hsu H-J, Bartenschlager R, Fischer WB. 2014. Membrane undulation induced by NS4A of Dengue virus: a molecular dynamics simulation study. *Journal of Biomolecular Structure and Dynamics* 32:1552–1562.
48. El Sahili A, Lescar J. 2017. Dengue Virus Non-Structural Protein 5. *Viruses* 9.
49. Zhu W, Chen CZ, Gorshkov K, Xu M, Lo DC, Zheng W. 2020. RNA-Dependent RNA Polymerase as a Target for COVID-19 Drug Discovery. *SLAS DISCOVERY: Advancing the Science of Drug Discovery* 247255522094212.
50. Ashour J, Laurent-Rolle M, Shi P-Y, García-Sastre A. 2009. NS5 of Dengue Virus Mediates STAT2 Binding and Degradation. *J Virol* 83:5408–5418.
51. Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* 20:1160–1166.
52. Friedman JH. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29:1189–1232.
53. Schapire RE. 2003. The Boosting Approach to Machine Learning: An Overview, p. 149–171. *In* Denison, DD, Hansen, MH, Holmes, CC, Mallick, B, Yu, B (eds.), *Nonlinear Estimation and Classification*. Springer New York, New York, NY.
54. Breiman L. 2001. Random forests. *Machine learning* 45:5–32.
55. Morais-Rodrigues F, Silvério-Machado R, Kato RB, Rodrigues DLN, Valdez-Baez J, Fonseca V, San EJ, Gomes LGR, dos Santos RG, Vinicius Canário Viana M, da Cruz Ferraz Dutra J, Teixeira Dornelles Parise M, Parise D, Campos FF, de Souza SJ, Ortega JM, Barh D, Ghosh P, Azevedo VAC, dos Santos MA. 2020. Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression. *Gene* 726:144168.
56. Goodsell DS, Zardecki C, Di Costanzo L, Duarte JM, Hudson BP, Persikova I, Segura J, Shao C, Voigt M, Westbrook JD, Young JY, Burley SK. 2020. RCSB Protein Data Bank: Enabling biomedical research and drug discovery. *Protein Sci* 29:52–65.
57. Schwede T, Kopp J, Guex N, Peitsch MC. 2003. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res* 31:3381–3385.

3.2 Artigo 2: “Machine learning and in-silico structural analysis as a tool for the rapid identification and characterization of circulating DENV lineages during Brazilian epidemics”

O Brasil passou por uma grande epidemia de vírus da dengue (DENV) em 2019, evidenciando dificuldades contínuas no seu controle efetivo e no preparo da saúde pública. A análise filogenética efetuada por (ADELINO et al., 2021) revelou padrões complexos de transmissão, com co-circulação e substituição de linhagens. Foram identificadas duas linhagens cocirculantes dentro do clado DENV2 BR-4. A metodologia de aprendizado de máquina desenvolvida para o trabalho anterior foi aplicada estas sequências, realizando a classificação entre as duas linhagens, na busca de assinaturas genéticas que as diferenciem. Foram identificadas 34 variações nucleotídicas únicas (SNVs) que caracterizam significativamente as linhagens BR-4L1 e BR-4L2, dentre as quais três resultam em substituição não sinônima de aminoácidos, A447V (ENV), K719I (NS5) e V553I (NS5). O presente trabalho ressalta a versatilidade e robustez da metodologia desenvolvida, aplicada a diferentes datasets e em diferentes perguntas científicas, o que corrobora seu uso como ferramenta complementar ao rol de análises realizadas em vigilância genômica em tempo real.

"Machine learning and in-silico structural analysis as a tool for the rapid identification and characterization of circulating DENV lineages during Brazilian epidemics"

Álvaro Salgado^a, Stephane Tosta^a, Marta Giovanetti^b, Vasco Azevedo^a, Jose Lourenço^c, Luiz Carlos J. Alcantara^b

^a Laboratório de Genética Celular e Molecular, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

^b Laboratório de Flavivírus, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil

^c Department of Zoology, University of Oxford, Oxford OX1 3PS, UK

Running Head: Machine learning applied to DENV genomic investigation

#Corresponding authors: Álvaro Salgado, alvarosalgado@ufmg.br; Luiz Carlos J Alcantara, luiz.alcantara@ioc.fiocruz.br; Jose Lourenço, jose.lourenco@zoo.ox.ac.uk

Abstract

Brazil experienced a major dengue virus (DENV) epidemic in 2019, highlighting ongoing difficulties in its effective disease control and public health preparation. The phylogenetic analysis performed by (ADELINO et al., 2021) revealed complex patterns of transmission, with co-circulation and lineage substitution. Two cocirculating lineages were identified within the DENV2 BR-4 clade. The machine learning methodology developed for the previous work was applied to these sequences, performing the classification between the two lineages, in search for genetic signatures that differentiate them. We identified 34 unique nucleotide variations (SNVs) that significantly characterize BR-4L1 and BR-4L2 strains, among which three result in non-synonymous amino acid substitutions, A447V (ENV), K719I (NS5) and V553I (NS5). The present work emphasizes the versatility and robustness of the methodology developed, applied to different datasets and for different scientific questions, which corroborates its use as a complementary tool to the ones used in real-time genomic surveillance.

Introduction

Dengue virus (DENV) is a global threat, posing health challenges to one third of the global human population, especially in locations where its vector, primarily mosquitoes of the genus *Aedes*, are most prevalent (KRAEMER et al., 2015). DENV infection causes a wide array of symptoms, ranging from asymptomatic self-limiting infections to severe hemorrhagic dengue fever (“Dengue; guidelines for diagnosis, treatment, prevention and control; rev. ed”, 2010).

DENV is a single-stranded, positive-sense RNA virus with a genome of approximately 11kb, belonging to the family Flaviviridae (genus *Flavivirus*). Its genome encodes a polyprotein comprising three structural (capsid, pre- membrane or membrane, and envelope) and seven non-structural (NS1, NS2a, NS2b, NS3, NS4a, NS4b, and NS5) proteins (RICE et al., 1985). It is classified into four serotypes (DENV1 – 4), each one encompassing different genotypes. Circulating viral diversity surveillance is of utmost importance for public health, as the most severe cases of DENV disease occur after serotype heterotypic infections (MARTINEZ et al., 2020).

Different serotypes have caused unexpectedly large epidemics in Brazil over the past 20 years, with the reemergence of DENV2 in the year of 2019 causing a serious epidemic, with 1,544,987 probable cases and 782 confirmed deaths (until epidemiological week – EW52 2019) (“Monitoramento dos casos de arboviroses urbanas transmitidas pelo *Aedes* (dengue, chikungunya e Zika), Semanas Epidemiológicas 01 a 52. 2020.”, [s.d.]).

Even though Brazil continues to fall short in its efforts to control mosquito vectors, the country has become a world reference in real-time genomic surveillance of arboviruses. In this context, in August 2019 a genomic surveillance training program, organized by PAHO and BrMoH, took place in Belo Horizonte (Minas Gerais state) under the title “Nanopore-based genome sequencing technology for temporal investigation and epidemiology of dengue outbreak: training, research, surveillance, and scientific dissemination” (LATIN AMERICAN GENOMIC SURVEILLANCE ARBOVIRAL NETWORK et al., 2021).

During this two-week training program, genomic sequencing and analysis of DENV samples from the 2019 epidemic were carried out. A total of 227 novel genomes were sequenced and analyzed, from which 170 sequences belonged do DENV2 serotype. Phylogenetic analyses revealed the presence of four different clades (DENV2 BR-1 – 4), with most of the sequences (164) belonging to BR-4 clade. Further analysis identified two distinct BR-4 lineages, named BR-4L1

and BR-4L2, that have established themselves and co-circulated in several Brazilian regions during this period.

In this work, we then proceeded to investigate if there were significant genetic signatures that differentiate these two lineages and, if so, what possible biological implications they could have.

Using machine learning techniques previously developed, we found 34 single nucleotide variants (SNVs) that significantly differentiate between BR-4L1 and BR4-L2. Three of those SNVs result in amino acid variation (non-synonymous variations), A447V (ENV protein), K719I (NS5 protein) and V553I (NS5 protein).

These non-synonymous variations were analyzed in silico by means of protein structural simulations.

The present work highlights the application of machine learning analyses as a fast and inexpensive first line of investigation in the context of real-time genomic surveillance.

Materials and methods

All sequences were obtained from public databases and contain primarily those generated during the two-week training program “Nanopore-based genome sequencing technology for temporal investigation and epidemiology of dengue outbreak: training, research, surveillance, and scientific dissemination” (LATIN AMERICAN GENOMIC SURVEILLANCE ARBOVIRAL NETWORK et al., 2021).

For machine learning analysis, nucleotide position numbering was determined by aligning the dataset to a reference DENV2 genome with known numbering and correct frame shift (NC_001474.2).

All sequences with coverage lower than 85% were removed from the dataset. Next, all nucleotide positions in which there was non-sequenced nucleotide (gap or “-“) were removed.

Then, all nucleotide positions that did not have variation between the sequences (uninformative positions) were removed. Next, one-hot-encoding was applied to binarize categorical attributes (i.e., nucleotides).

Three different machine learning algorithms were used to classify between BR-4L1 and BR-4L2 sequences, XGBoost (XGB) (FRIEDMAN, 2001; SCHAPIRE, 2003), Random Forest (RF) (BREIMAN, 2001) and Regularized Logistic Regression (RLG) (MORAIS-RODRIGUES et al., 2020), aggregating their results. The dataset was split between training and testing (50%), and hyperparameters were adjusted with cross-validation.

The adjusted models (XGB and RF) were interpreted with SHAP algorithm (LUNDBERG et al., 2018; LUNDBERG; ERION; LEE, 2018; LUNDBERG; LEE, 2017) to identify nucleotide positions most associated with the classification results. For RLG, the parameters with the highest values were chosen as the ones most associated with the classification results.

Results

All three ML models adopted were able to classify between BR-4L1 and BR-4L2 sequences with 100% accuracy when applied to the test dataset

The nucleotide (NN) positions whose variations are responsible for discriminating between the two lineages are shown in the table below. The table also shows in which protein the SNVs occur, the amino acid position on the protein, as well as the codon and translated amino acid. As mentioned earlier, three SNVs resulted in amino acid variation (highlighted on the table). It also shows which codon and amino acid were present on most recent common ancestor (MRCA) to both lineages.

The last column shows the Shannon entropy for that nucleotide for each lineage (BR-4L1 and BR-4L2). A value of 0 (zero) means all samples from that lineage possess the same nucleotide in that position. It is important to notice that this implies a perfect distinction between the two lineages based on those nucleotides whose entropies are 0.00/0,00.

Table 3 – Machine Learning analysis results

NN position	Protein	AA position on protein	MRCA	BR-4L1	BR-4L2	codon MRCA	codon BR-4L1	codon BR-4L2	Entropy (of nn position within lineage)
450	membrane glycoprotein precursor prM	4	T	T	T	acc	act	acc	0.00 / 0.00
954	envelope protein E	6	I	I	I	ata	atc	ata	0.00 / 0.00
2276	envelope protein E	447	A	A	V	gct	gct	gtt	0.00 / 0.00
2319	envelope protein E	461	V	V	V	gtc	gta	gtc	0.00 / 0.00
2448	nonstructural protein NS1	9	K	K	K	aaa	aag	aaa	0.00 / 0.00
3330	nonstructural protein NS1	303	A	A	A	gcc	gct	gcc	0.00 / 0.00
5304	nonstructural protein NS3	261	C	C	C	tgt	tgc	tgt	0.00 / 0.12
5730	nonstructural protein NS3	403	D	D	D	gac	gac	gat	0.00 / 0.12
5190	nonstructural protein NS3	223	P	P	P	ccc	cct	ccc	0.00 / 0.00
7059	nonstructural protein NS4B	78	L	L	L	ctt	ctc	ctt	0.00 / 0.00
4482	nonstructural protein NS2B	117	T	T	T	acg	aca	acg	0.00 / 0.00
7038	nonstructural protein NS4B	71	Q	Q	Q	caa	caa	cag	0.00 / 0.00
9399	RNA-dependent RNA polymerase NS5	610	N	N	N	aat	aac	aat	0.00 / 0.00
7758	RNA-dependent RNA polymerase NS5	63	R	R	R	aga	aga	agg	0.00 / 0.00
9795	RNA-dependent RNA polymerase NS5	742	Q	Q	Q	cag	caa	cag	0.00 / 0.00
5835	nonstructural protein NS3	438	E	E	E	gaa	gag	gaa	0.00 / 0.00
9226	RNA-dependent RNA polymerase NS5	553	V	I	V	gta	ata	gta	0.00 / 0.00
9942	RNA-dependent RNA polymerase NS5	791	S	S	S	agt	agt	agc	0.00 / 0.00
8298	RNA-dependent RNA polymerase NS5	243	F	F	F	ttc	ttt	ttc	0.00 / 0.00
7999	RNA-dependent RNA polymerase NS5	144	L	L	L	ctg	ttg	ctg	0.00 / 0.00
9208	RNA-dependent RNA polymerase NS5	547	L	L	L	tta	cta	tta	0.00 / 0.00
2223	envelope protein E	429	F	F	F	ttt	ttc	ttt	0.00 / 0.12
8658	RNA-dependent RNA polymerase NS5	363	T	T	T	act	acc	act	0.10 / 0.00
9424	RNA-dependent RNA polymerase NS5	619	L	L	L	tta	cta	tta	0.00 / 0.12
7137	nonstructural protein NS4B	104	P	P	P	cct	ccc	cct	0.41 / 0.12
9576	RNA-dependent RNA polymerase NS5	669	P	P	P	cct	ccc	cct	0.30 / 0.12
9660	RNA-dependent RNA polymerase NS5	697	S	S	S	tca	tca	tct	0.00 / 1.00
4341	nonstructural protein NS2B	70	S	S	S	agt	agc	agt	0.00 / 0.97
6966	nonstructural protein NS4B	47	F	F	F	ttt	ttt	ttt	0.00 / 1.00
9725	RNA-dependent RNA polymerase NS5	719	K	K	I	aaa	aaa	ata	0.00 / 1.00
1986	envelope protein E	350	R	R	R	cgc	cgc	cgc	0.00 / 1.00
9303	RNA-dependent RNA polymerase NS5	578	V	V	V	gtg	gtg	gtg	0.00 / 1.00
7656	RNA-dependent RNA polymerase NS5	29	K	K	K	aag	aag	aag	0.18 / 1.00
4974	nonstructural protein NS3	151	G	G	G	ggc	ggc	ggg	0.73 / 0.69

Discussion

In this study, we demonstrate the potential of applying ML approaches on real-time genomic surveillance, to quickly identify genetic loci which may be of public health interest. The same framework developed for a previous work, involving aggregating machine learning models and their interpretation using SHAP, was used on this work, which demonstrate its versatility and robustness and ease of application to new datasets and different scientific questions.

It is interesting to note the high accuracy of the machine learning models in classifying the data. This was expected, since previous phylogenetic analysis has grouped the sequences in two well defined lineages. The fact that the machine learning models perfectly differentiate between the two lineages is in accordance with the biological process being investigated and corroborates the method's applicability.

A447V and K719I variations were identified in the BR-4L1, while V553I was identified in BR-4L2. A447V (ENV protein) and V553I (NS5 protein) appear to be conservative changes due to the interchangeable character for the respective amino acids. In contrast, K719I in the NS5 protein has changed from a negatively charged to an aliphatic amino acid.

Further investigation involving *in-silico* protein analysis is necessary to suggest possible impacts of those aminoacid substitutions on protein function. Also, selection pressure analysis could inform on whether these substitutions are becoming fixed on the population or not.

Conclusion

In conclusion, our study uses efficient and fast computational approaches to analyzing sequence data, making it suitable for real-time genomic surveillance, in which genetic data is analyzed as it is generated. This approach may help detect and inform on possible connections between ongoing genetic changes and public health in a timely manner.\

Brazil has become a global reference in real-time genomic surveillance. The methods developed and applied in this work can integrate the list of analytic tools used for such, giving complimentary results and helping obtain the most information from data available.

References

- ADELINO, T. É. R. et al. Field and classroom initiatives for portable sequence-based monitoring of dengue virus in Brazil. **Nature communications**, v. 12, n. 1, p. 1–12, 2021.
- BACH, S. et al. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. **PLOS ONE**, v. 10, n. 7, p. e0130140, 10 jul. 2015.
- BAO, L.; ZHOU, M.; CUI, Y. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. **Nucleic Acids Research**, v. 33, n. suppl_2, p. W480–W482, 1 jul. 2005.
- BARROWS, N. J. et al. Biochemistry and Molecular Biology of Flaviviruses. **Chemical Reviews**, v. 118, n. 8, p. 4448–4482, 25 abr. 2018.
- BEHRAVAN, H. et al. Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls. **Scientific Reports**, v. 8, n. 1, dez. 2018.
- BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5–32, 2001.
- BRYANT, J. E.; HOLMES, E. C.; BARRETT, A. D. T. Out of Africa: a molecular perspective on the introduction of yellow fever virus into the Americas. **PLoS pathogens**, v. 3, n. 5, p. e75, 2007.
- CHAMBERS, T. J. et al. Flavivirus genome organization, expression, and replication. **Annual Review of Microbiology**, v. 44, n. 1, p. 649–688, 1 out. 1990.
- CHEN, S. et al. Innate immune evasion mediated by flaviviridae non-structural proteins. **Viruses**, v. 9, n. 10, p. 291, 2017.
- CHEN, T.; GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System**. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais... Em: KDD '16: THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING**. San Francisco California USA: ACM, 13 ago. 2016. Disponível em: <<https://dl.acm.org/doi/10.1145/2939672.2939785>>. Acesso em: 1 maio. 2020
- CHEN, X. et al. A forest-based approach to identifying gene and gene–gene interactions. **Proceedings of the National Academy of Sciences**, v. 104, n. 49, p. 19199–19203, 4 dez. 2007.
- CHEN, X.; ISHWARAN, H. Random forests for genomic data analysis. **Genomics**, v. 99, n. 6, p. 323–329, 1 jun. 2012.
- CLETON, N. et al. Come fly with me: review of clinically important arboviruses for global travelers. **Journal of Clinical Virology**, v. 55, n. 3, p. 191–203, 2012.
- CONSOLI, R. A.; DE OLIVEIRA, R. L. **Principais mosquitos de importância sanitária no Brasil**. [s.l.] SciELO-Editora FIOCRUZ, 1994.

DATTA, A.; SEN, S.; ZICK, Y. **Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems**. 2016 IEEE Symposium on Security and Privacy (SP). **Anais...** Em: 2016 IEEE SYMPOSIUM ON SECURITY AND PRIVACY (SP). San Jose, CA: IEEE, maio 2016. Disponível em: <<http://ieeexplore.ieee.org/document/7546525/>>. Acesso em: 23 jul. 2019

Dengue; guidelines for diagnosis, treatment, prevention and control; rev. ed. **SciTech Book News**, jun. 2010.

DOMINGOS, P. **The master algorithm: how the quest for the ultimate learning machine will remake our world**. New York: Basic Books, a member of the Perseus Books Group, 2015.

FARIA, N. R. et al. Mobile real-time surveillance of Zika virus in Brazil. **Genome Medicine**, v. 8, n. 1, p. 97, 29 set. 2016.

FARIA, N. R. et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. **Nature**, v. 546, n. 7658, p. 406–410, jun. 2017.

FIGUEIREDO, L. T. M. Human Urban Arboviruses Can Infect Wild Animals and Jump to Sylvatic Maintenance Cycles in South America. **Frontiers in Cellular and Infection Microbiology**, v. 9, p. 259, 17 jul. 2019.

FRIEDMAN, J. H. Greedy Function Approximation: A Gradient Boosting Machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189–1232, 2001.

GARDNER, C. L.; RYMAN, K. D. Yellow Fever: A Reemerging Threat. **Clinics in Laboratory Medicine**, v. 30, n. 1, p. 237–260, mar. 2010.

HO, D. S. W. et al. Machine Learning SNP Based Prediction for Precision Medicine. **Frontiers in Genetics**, v. 10, 27 mar. 2019.

HOLBROOK, M. R. Historical perspectives on flavivirus research. **Viruses**, v. 9, n. 5, p. 97, 2017.

ISHWARAN, H. et al. High-Dimensional Variable Selection for Survival Data. **Journal of the American Statistical Association**, v. 105, n. 489, p. 205–217, 1 mar. 2010.

JONES, K. E. et al. Global trends in emerging infectious diseases. **Nature**, v. 451, n. 7181, p. 990–993, 2008.

KRAEMER, M. U. G. et al. The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*. **eLife**, v. 4, p. e08347–e08347, 2015.

LATIN AMERICAN GENOMIC SURVEILLANCE ARBOVIRAL NETWORK et al. Field and classroom initiatives for portable sequence-based monitoring of dengue virus in Brazil. **Nature Communications**, v. 12, n. 1, p. 2296, abr. 2021.

LI, P. An Empirical Evaluation of Four Algorithms for Multi-Class Classification: Mart, ABC-Mart, Robust LogitBoost, and ABC-LogitBoost. **arXiv:1001.1020 [cs]**, 7 jan. 2010.

LIPOVETSKY, S.; CONKLIN, M. Analysis of regression in game theory approach. **Applied Stochastic Models in Business and Industry**, v. 17, n. 4, p. 319–330, out. 2001.

LOPES, N.; NOZAWA, C.; LINHARES, R. E. C. Características gerais e epidemiologia dos arbovírus emergentes no Brasil. **Revista Pan-Amazônica de Saúde**, v. 5, n. 3, p. 10–10, 2014.

LUBKE, G. et al. Gradient Boosting as a SNP Filter: an Evaluation Using Simulated and Hair Morphology Data. **Journal of data mining in genomics & proteomics**, v. 4, 20 out. 2013.

LUNDBERG, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. **Nature Biomedical Engineering**, v. 2, n. 10, p. 749–760, out. 2018.

LUNDBERG, S. M.; ERION, G. G.; LEE, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. **arXiv:1802.03888 [cs, stat]**, 11 fev. 2018.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. Em: GUYON, I. et al. (Eds.). **Advances in neural information processing systems 30**. [s.l.] Curran Associates, Inc., 2017. p. 4765–4774.

LUNDBERG, S. M.; LEE, S.-I. Consistent feature attribution for tree ensembles. **arXiv:1706.06060 [cs, stat]**, 16 fev. 2018.

MARES-GUIA, M. A. M. DE M. et al. Yellow fever epizootics in non-human primates, Southeast and Northeast Brazil (2017 and 2018). **Parasites & Vectors**, v. 13, n. 1, 2020.

MARTINEZ, D. R. et al. Antigenic variation of the dengue virus 2 genotypes impacts the neutralization activity of human antibodies in vaccinees. **Cell reports**, v. 33, n. 1, p. 108226, 2020.

MINISTÉRIO DA SAÚDE DO BRASIL. **Vigilância em saúde no Brasil 2003|2019: da criação da Secretaria de Vigilância em Saúde aos dias atuais**. , 2019. Disponível em: <<https://www.saude.gov.br/images/pdf/2020/janeiro/20/Boletim-epidemiologico-SVS-02-1-.pdf>>

MINISTÉRIO DA SAÚDE DO BRASIL. **Monitoramento dos casos de arboviroses urbanas transmitidas pelo Aedes (dengue, chikungunya e Zika), Semanas Epidemiológicas 01 a 52. 2020**. , 2020. Disponível em: <<https://www.saude.gov.br/images/pdf/2020/janeiro/20/Boletim-epidemiologico-SVS-02-1-.pdf>>

MONATH, T. P. Yellow fever: an update. **The Lancet. Infectious Diseases**, v. 1, n. 1, p. 11–20, ago. 2001.

MONATH, T. P. Treatment of yellow fever. **Antiviral Research**, v. 78, n. 1, p. 116–124, abr. 2008.

Monitoramento dos casos de arboviroses urbanas transmitidas pelo Aedes (dengue, chikungunya e Zika), Semanas Epidemiológicas 01 a 52. 2020. Brasil. Ministério da Saúde., , [s.d.]. Disponível em: <<https://www.saude.gov.br/images/pdf/2020/janeiro/20/Boletim-epidemiologico-SVS-02-1-.pdf>>. Acesso em: 20 maio. 2020

MOORE, J. H.; ASSELBERGS, F. W.; WILLIAMS, S. M. Bioinformatics challenges for genome-wide association studies. **Bioinformatics**, v. 26, n. 4, p. 445–455, 15 fev. 2010.

MORAIS-RODRIGUES, F. **Microarray-based breast cancer classification using logistic regression and beyond**. [s.l.] Universidade Federal de Minas Gerais, 2017.

MORAIS-RODRIGUES, F. et al. Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression. **Gene**, v. 726, p. 144168, fev. 2020.

MORELI, M. L.; COSTA, V. A systematic review of molecular diagnostic methods for the detection of arboviruses in clinical specimens in Brazil and the importance of a differential diagnosis. **Virology Discovery**, v. 1, n. 1, p. 1, 2013.

MUKTAR, Y.; TAMERAT, N.; SHEWAFERA, A. Aedes aegypti as a Vector of Flavivirus. **J Trop Dis**, v. 4, n. 223, p. 2, 2016.

MUSSO, D.; GUBLER, D. J. Zika virus. **Clinical microbiology reviews**, v. 29, n. 3, p. 487–524, 2016.

PAHO/WHO. **Laboratory Diagnosis of Yellow Fever Virus infection - 2018 - PAHO/WHO | Pan American Health Organization**. Disponível em: <<https://www.paho.org/en/documents/laboratory-diagnosis-yellow-fever-virus-infection-2018>>. Acesso em: 23 jan. 2022.

PAHO/WHO. **Dengue - PAHO/WHO | Pan American Health Organization**. Disponível em: <<https://www.paho.org/en/topics/dengue>>. Acesso em: 24 jan. 2022a.

PAHO/WHO. **Epidemiological Update: Yellow Fever - 28 December 2021 - PAHO/WHO | Pan American Health Organization**. Disponível em: <<https://www.paho.org/en/documents/epidemiological-update-yellow-fever-28-december-2021>>. Acesso em: 23 jan. 2022b.

PAHO/WHO. **Yellow fever**. Disponível em: <<https://www.paho.org/en/topics/yellow-fever>>. Acesso em: 23 jan. 2022.

PAULES, C. I.; FAUCI, A. S. Yellow Fever — Once Again on the Radar Screen in the Americas. **New England Journal of Medicine**, v. 376, n. 15, p. 1397–1399, 13 abr. 2017.

PIERSON, T. C.; DIAMOND, M. S. The continued threat of emerging flaviviruses. **Nature microbiology**, v. 5, n. 6, p. 796–812, 2020.

QI, Y. Random forest for bioinformatics. Em: **Ensemble machine learning**. [s.l.] Springer, 2012. p. 307–323.

RIBEIRO, M. T.; SINGH, S.; GUESTIN, C. “**Why Should I Trust You?**”: **Explaining the Predictions of Any Classifier**. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. **Anais...** Em: THE 22ND ACM SIGKDD

INTERNATIONAL CONFERENCE. San Francisco, California, USA: ACM Press, 2016. Disponível em: <<http://dl.acm.org/citation.cfm?doid=2939672.2939778>>. Acesso em: 23 jul. 2019

RICE, C. M. et al. Nucleotide Sequence of Yellow Fever Virus: Implications for Flavivirus Gene Expression and Evolution. **Science (American Association for the Advancement of Science)**, v. 229, n. 4715, p. 726–733, 1985.

SAN MARTÍN, J. L. et al. The epidemiology of dengue in the Americas over the last three decades: a worrisome reality. **The American journal of tropical medicine and hygiene**, v. 82, n. 1, p. 128, 2010.

SCHAPIRE, R. E. The Boosting Approach to Machine Learning: An Overview. Em: DENISON, D. D. et al. (Eds.). **Nonlinear Estimation and Classification**. Lecture Notes in Statistics. New York, NY: Springer New York, 2003. v. 171p. 149–171.

SHEARER, F. M. et al. Global yellow fever vaccination coverage from 1970 to 2016: an adjusted retrospective analysis. **The Lancet infectious diseases**, v. 17, n. 11, p. 1209–1217, 2017.

SHRIKUMAR, A.; GREENSIDE, P.; KUNDAJE, A. Learning Important Features Through Propagating Activation Differences. **arXiv:1704.02685 [cs]**, 9 abr. 2017.

SILVA, N. I. O. et al. Recent sylvatic yellow fever virus transmission in Brazil: the news from an old disease. **Virology Journal**, v. 17, n. 1, p. 9, 23 jan. 2020.

SIMMONDS, P. et al. ICTV virus taxonomy profile: Flaviviridae. **The Journal of general virology**, v. 98, n. 1, p. 2, 2017.

ŠTRUMBELJ, E.; KONONENKO, I. Explaining prediction models and individual predictions with feature contributions. **Knowledge and Information Systems**, v. 41, n. 3, p. 647–665, dez. 2014.

VASCONCELOS, P. F. DA C. Yellow fever in Brazil: thoughts and hypotheses on the emergence in previously free areas. **Revista de Saúde Pública**, v. 44, n. 6, p. 1144–1149, dez. 2010.

VASCONCELOS, P. F.; MONATH, T. P. Yellow fever remains a potential threat to public health. **Vector-Borne and zoonotic diseases**, v. 16, n. 8, p. 566–567, 2016.

WANG, W. Y. S. et al. Genome-wide association studies: theoretical and practical concerns. **Nature Reviews Genetics**, v. 6, n. 2, p. 109, fev. 2005.

WHO. **Yellow fever**. Disponível em: <<https://www.who.int/news-room/fact-sheets/detail/yellow-fever>>. Acesso em: 23 jan. 2022.

WHO SCIENTIFIC GROUP ON ARBOVIRUSES. **Arboviruses and human disease : report of a WHO scientific group [meeting held in Geneva from 26 September to 1 October 1966].** : World Health Organization technical report series ; no. 369. World Health Organization, , 1967.

WOLPERT, D. H. The Lack of A Priori Distinctions Between Learning Algorithms. **Neural Computation**, v. 8, n. 7, p. 1341–1390, out. 1996.

WU, P. et al. Arbovirus lifecycle in mosquito: acquisition, propagation and transmission. **Expert reviews in molecular medicine**, v. 21, 2019.

4 Discussão geral

Vírus emergentes e reemergentes apresentam um desafio de grande complexidade para o sistema de saúde pública brasileiro. Dentre eles, as arboviroses transmitidas por mosquitos são agentes capazes de causar doenças graves, como febres hemorrágicas, encefalites e meningites. Por esses motivos, a vigilância genômica em tempo real é de extrema importância para guiar ações de controle e prevenção, pois permite a reconstrução das origens da epidemia e a estimativa das taxas de transmissão em diferentes momentos e regiões geográficas, sujeitas a fatores ambientais e humanos. Além disso, a vigilância genômica torna possível a identificação de variantes emergentes, reemergentes, circulantes e cocirculantes, por meio da quantificação da diversidade genética viral, possibilitando alertar sobre a probabilidade de novos surtos e/ou possível escape a vacinas e tratamentos existentes. Como resultado, adquirem-se informações relevantes para o delineamento de políticas públicas de saúde, além da contribuição com o desenvolvimento de vacinas, novos medicamentos, aperfeiçoamento dos métodos diagnósticos sorológicos e moleculares.

Nesse contexto, o Brasil vem se tornando referência global em vigilância genômica em tempo real, atingindo resultados fundamentais na detecção precoce e acompanhamento de surtos. No entanto, a grande quantidade de dados produzidos com a utilização de plataformas de sequenciamento de nova geração demanda abordagens analíticas sofisticadas, capazes de lidar com dados complexos e volumosos, que possibilitem a extração da maior quantidade de informação possível. Nesse sentido, algoritmos de Aprendizado de Máquina têm sido empregados com sucesso em Bioinformática, motivando sua aplicação na busca por assinaturas genéticas em arbovírus, associadas a características fenotípicas ou epidemiológicas em surtos recentes no Brasil.

Neste trabalho, foi desenvolvida uma estrutura analítica baseada em aprendizado de máquina, aplicada a sequências genômicas de arbovírus, com o objetivo de identificar assinaturas genéticas que pudessem estar relacionadas a observações clínicas, epidemiológicas e filogenéticas de interesse. Para tanto, aliou-se três algoritmos robustos e versáteis, Random Forest, XGBoost e Regressão Logística Regularizada, ao método de interpretação de modelos SHAP, o que possibilitou obter resultados consistentes com os processos biológicos subjacentes.

Dentre os principais surtos recentes de arboviroses no Brasil, a febre amarela figura como caso que inspira preocupação, devido à severidade da doença causada pela infecção e sua expansão para regiões além daquelas consideradas endêmicas até então, trazendo o temor de estabelecimento de ciclo urbano de transmissão em regiões densamente povoadas, onde a cobertura vacinal é relativamente baixa.

A ameaça de estabelecimento de ciclo urbano de transmissão é reforçada pelo padrão observado nas amostras de YFV provenientes de primatas não humanos, em especial aqueles do gênero *Callithrix*, para os quais observou-se uma variação grande nos valores Ct. Um grande número de amostras de *Callithrix* apresentava valores altos quando comparados a outros gêneros, como *Alouatta*. Esta observação sugere a possibilidade de que tais animais, altamente prevalentes em regiões urbanas e periurbanas, como nos estados das regiões sul e sudeste do Brasil, possam apresentar baixa viremia durante a infecção por YFV e persistir por mais tempo, prolongando o risco de transmissão entre animais, com potencial de spillover para humanos.

Além disso, a alta taxa de mortalidade da febre amarela em humanos é preocupante, de modo que propôs-se investigar possíveis associações entre variantes genéticas do vírus à severidade da doença, de acordo com informações clínicas relacionadas às sequências.

Portanto, para as sequências provenientes de infecções humanas, foram buscados padrões relacionados à severidade da doença. Para aquelas provenientes de primatas não humanos, investigou-se a possível relação com o valor Ct relatado nas amostras. Assim, o trabalho identificou sinais em diversos nucleotídeos, tendo sido selecionados aqueles que resultaram em variação não sinônima do respectivo aminoácido. Foram realizadas, então, análises estruturais das proteínas onde ocorreram essas variações.

Na proteína do envelope (E) foi identificada uma variação L395F, com possível alteração na plasticidade de uma região importante para a adesão e fusão viral à célula do hospedeiro, em sequências provenientes de primatas não humanos. Ainda em sequências provenientes de *Callithrix*, há uma variação A171V na proteína não estrutural NS1, que tem papel importante na interação do vírus com o sistema imune do hospedeiro, além de modular a permeabilidade capilar e a severidade de infecções. Nas sequências provenientes de humanos, foi identificada a variação E614D na proteína não estrutural NS3, cujas análises in-silico adicionais indicaram possível redução na afinidade por RNA e impacto na função da proteína, replicação do genoma viral. Ainda

nas sequências provenientes de humanos, uma das variações presentes na proteína não estrutural NS5 pode ter impacto na função de polimerase de RNA dependente de RNA, com consequente alteração na capacidade de replicação viral.

As hipóteses sugeridas acima servem como ponto inicial para investigações subsequentes, tanto com um número maior de sequências quanto com metodologias de bancada.

Outro agente infeccioso que causa preocupação constante à população brasileira é o vírus da dengue. Apesar de ter sido erradicado de várias regiões da América do Sul em meados do século XX, ressurgiu no Brasil na década de 80 e, desde então, vem causando surtos nas regiões mais ao sul do país, com reemergência de diferentes sorotipos ao longo das décadas. O surto mais recente apresentou predominância dos sorotipos DENV1 e DENV2, para os quais foram gerados um total de 227 novas sequências genômicas completas, das quais mais de 75% foram processadas e analisadas durante o programa de treinamento e ensino em sequenciamento baseado em Nanoporos, realizado em 2019 na cidade de Belo Horizonte, Minas Gerais.

As sequências geradas correspondem a mais da metade daquelas atualmente disponíveis em bancos de dados públicos, o que ressalta o grande feito do projeto mas também a carência atual em sequências disponíveis de vírus da dengue. Essa carência de sequências pertencentes a surtos anteriores impactou negativamente na capacidade do grupo em aprofundar algumas análises filogenéticas e epidemiológicas, como aquelas relacionadas a eventos de substituição de linhagem circulante. Mesmo assim, foi possível identificar a cocirculação de diferentes linhagens durante o surto de 2019.

No que diz respeito aos casos clínicos, sabe-se que a infecção por sorotipos diferentes está associada a uma maior severidade da doença. Porém, apesar do aumento no número absoluto de casos e óbitos em 2019, a reemergência do sorotipo DENV2 nas regiões afetadas não apresentou associação com aumento significativo na taxa de mortalidade quando comparada aos anos anteriores.

Foi possível determinar que todas as sequências do sorotipo DENV2 pertenciam ao genótipo III, sendo agrupados em quatro clados (BR-1 a BR-4). Foi possível identificar que o clado BR-4, mais recente, circulou na forma de duas linhagens (BR-4L1 e BR-4L2). Foi aplicada, então, a metodologia baseada em Aprendizado de Máquina para discriminar sequências das duas linhagens, buscando destacar variações de nucleotídeo que sejam características de cada e as

diferenciem. Foram observadas três variações nucleotídicas que resultaram em modificação de aminoácidos. Destas, duas ocorrem em todas as sequências da linhagem correspondente, ressaltando a concordância desses achados com as análises filogenéticas e fortalecendo a validade biológica da metodologia empregada. Considerando que a análise filogenética, por si só, não destaca individualmente essas variações de nucleotídeo único, os resultados reforçam a importância de aliar a metodologia desenvolvida neste trabalho às análises filogenéticas tradicionalmente aplicadas em vigilância genômica em tempo real, visto que torna possível identificar precisamente quais variações genéticas são características de diferentes linhagens em circulação e das suas propriedades fenotípicas de interesse de saúde pública, informando sobre possíveis implicações biológicas e intervenções.

Análises adicionais, como modelagem estrutural de proteínas afetadas por essas variantes genéticas, são necessárias para propor possíveis impactos no fitness e patogênese viral.

Apesar dos resultados positivos nos artigos descritos na seção de resultados, o número limitado de sequências analisadas exige cautela ao apresentar conclusões, pois análises baseadas em aprendizado de máquina se beneficiam de dados de alta qualidade e em grande quantidade. Isso é especialmente importante ao lidar com dados genômicos de vírus, cuja alta taxa de mutação pode inserir variações que não têm impacto biológico significativo, configurando ruído que só pode ser devidamente filtrado com a análise de um número grande de sequências. Além disso, não levamos em consideração as interações vírus-hospedeiro, que certamente possuem influência sobre a dinâmica da doença, sobretudo na análise da severidade de infecções humanas. Nesse sentido, é importante priorizar a adequada coleta e subsequente disponibilização de dados clínicos, atrelados às amostras e sequências virais.

5 Conclusão

A vigilância genômica em tempo real gera dados valiosos para a identificação precoce de surtos e emergência ou reemergência de patógenos. Auxiliam sobremaneira no acompanhamento da história evolutiva, geográfica e temporal de vírus, ajudando a elucidar possíveis conexões entre variações genéticas em andamento e seu impacto em questões de saúde pública, o que possibilita uma resposta ágil e direcionada aos desafios apresentados por agentes infecciosos.

O trabalho aqui apresentado elaborou uma metodologia de análise genômica em tempo real baseada em Aprendizado de Máquina, de aplicação rápida e eficiente, facilmente adaptável para diferentes estudos, nos quais os dados são analisados à medida em que são gerados. As metodologias desenvolvidas utilizam uma ferramenta de interpretação de modelos baseada em teoria dos jogos, constituindo uma abordagem não enviesada, rápida, barata, robusta.

Diferentes fenômenos biológicos de interesse foram investigados com o uso da mesma metodologia, como variações de valor Ct em sequências de YFV de primatas não humanos, severidade da doença em infecções por YFV em humanos e características discriminantes de linhagens cocirculantes de DENV2 definidas por filogenia. Isso corrobora a versatilidade do método desenvolvido e sua capacidade de auxiliar em estudos, realizados sob diferentes ângulos, em um mesmo conjunto de dados.

É importante observar a contribuição da metodologia proposta ao contexto de vigilância genômica em tempo real. Os dados genômicos analisados contém uma diversidade grande de variações nucleotídicas, que são analisadas na análise filogenética para descrever a evolução espaço-temporal dos surtos e epidemias. Porém, assinaturas específicas de fenômenos biológicos distintos (fenótipos) não podem ser determinadas somente com análise filogenética. Dessa forma, a utilização da metodologia proposta permite identificar essas assinaturas de interesse, dentre todas as variações nucleotídicas presentes no banco de dados.

Por isso, o trabalho desenvolvido configura um estudo teste que demonstra e reforça a importância de se utilizar a metodologia baseada em aprendizado de máquina, conjuntamente com as demais ferramentas de análise empregadas em vigilância genômica em tempo real, sendo interessante incluí-la como parte integrante do rol de ferramentas analíticas no futuro.

Entretanto, as assinaturas genéticas identificadas, isoladamente, podem não ser suficientes para esclarecer os fundamentos biológicos por trás dos fenômenos observados. Para tanto, é necessário sempre aprofundar o estudo, seja por meio de outras análises de Bioinformática, como modelagem estrutural de proteínas, ou em experimentos de bancada.

Além disso, bons resultados dependem da disponibilidade de grande quantidade de sequências genômicas de qualidade, associadas a dados clínicos e epidemiológicos detalhados e confiáveis. Sendo assim, é necessário financiamento contínuo para os esforços de vigilância genômica em tempo real.

Concluindo, a metodologia desenvolvida serve como ferramental adicional disponível ao grupo de pesquisa, a ser utilizado e aperfeiçoado por futuros colaboradores e, a partir do qual, novos métodos de análise de dados baseados em Aprendizado de Máquina podem ser desenvolvidos.

Referências Bibliográficas

- ADELINO, T. É. R. et al. Field and classroom initiatives for portable sequence-based monitoring of dengue virus in Brazil. **Nature communications**, v. 12, n. 1, p. 1–12, 2021.
- BACH, S. et al. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. **PLOS ONE**, v. 10, n. 7, p. e0130140, 10 jul. 2015.
- BAO, L.; ZHOU, M.; CUI, Y. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. **Nucleic Acids Research**, v. 33, n. suppl_2, p. W480–W482, 1 jul. 2005.
- BARROWS, N. J. et al. Biochemistry and Molecular Biology of Flaviviruses. **Chemical Reviews**, v. 118, n. 8, p. 4448–4482, 25 abr. 2018.
- BEHRAVAN, H. et al. Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls. **Scientific Reports**, v. 8, n. 1, dez. 2018.
- BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5–32, 2001.
- BRYANT, J. E.; HOLMES, E. C.; BARRETT, A. D. T. Out of Africa: a molecular perspective on the introduction of yellow fever virus into the Americas. **PLoS pathogens**, v. 3, n. 5, p. e75, 2007.
- CHAMBERS, T. J. et al. Flavivirus genome organization, expression, and replication. **Annual Review of Microbiology**, v. 44, n. 1, p. 649–688, 1 out. 1990.
- CHEN, S. et al. Innate immune evasion mediated by flaviviridae non-structural proteins. **Viruses**, v. 9, n. 10, p. 291, 2017.
- CHEN, T.; GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System**. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais... Em: KDD '16: THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING**. San Francisco California USA: ACM, 13 ago. 2016. Disponível em: <<https://dl.acm.org/doi/10.1145/2939672.2939785>>. Acesso em: 1 maio. 2020
- CHEN, X. et al. A forest-based approach to identifying gene and gene–gene interactions. **Proceedings of the National Academy of Sciences**, v. 104, n. 49, p. 19199–19203, 4 dez. 2007.
- CHEN, X.; ISHWARAN, H. Random forests for genomic data analysis. **Genomics**, v. 99, n. 6, p. 323–329, 1 jun. 2012.
- CLETON, N. et al. Come fly with me: review of clinically important arboviruses for global travelers. **Journal of Clinical Virology**, v. 55, n. 3, p. 191–203, 2012.
- CONSOLI, R. A.; DE OLIVEIRA, R. L. **Principais mosquitos de importância sanitária no Brasil**. [s.l.] SciELO-Editora FIOCRUZ, 1994.

DATTA, A.; SEN, S.; ZICK, Y. **Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems**. 2016 IEEE Symposium on Security and Privacy (SP). **Anais...** Em: 2016 IEEE SYMPOSIUM ON SECURITY AND PRIVACY (SP). San Jose, CA: IEEE, maio 2016. Disponível em: <<http://ieeexplore.ieee.org/document/7546525/>>. Acesso em: 23 jul. 2019

Dengue; guidelines for diagnosis, treatment, prevention and control; rev. ed. **SciTech Book News**, jun. 2010.

DOMINGOS, P. **The master algorithm: how the quest for the ultimate learning machine will remake our world**. New York: Basic Books, a member of the Perseus Books Group, 2015.

FARIA, N. R. et al. Mobile real-time surveillance of Zika virus in Brazil. **Genome Medicine**, v. 8, n. 1, p. 97, 29 set. 2016.

FARIA, N. R. et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. **Nature**, v. 546, n. 7658, p. 406–410, jun. 2017.

FIGUEIREDO, L. T. M. Human Urban Arboviruses Can Infect Wild Animals and Jump to Sylvatic Maintenance Cycles in South America. **Frontiers in Cellular and Infection Microbiology**, v. 9, p. 259, 17 jul. 2019.

FRIEDMAN, J. H. Greedy Function Approximation: A Gradient Boosting Machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189–1232, 2001.

GARDNER, C. L.; RYMAN, K. D. Yellow Fever: A Reemerging Threat. **Clinics in Laboratory Medicine**, v. 30, n. 1, p. 237–260, mar. 2010.

HO, D. S. W. et al. Machine Learning SNP Based Prediction for Precision Medicine. **Frontiers in Genetics**, v. 10, 27 mar. 2019.

HOLBROOK, M. R. Historical perspectives on flavivirus research. **Viruses**, v. 9, n. 5, p. 97, 2017.

ISHWARAN, H. et al. High-Dimensional Variable Selection for Survival Data. **Journal of the American Statistical Association**, v. 105, n. 489, p. 205–217, 1 mar. 2010.

JONES, K. E. et al. Global trends in emerging infectious diseases. **Nature**, v. 451, n. 7181, p. 990–993, 2008.

KRAEMER, M. U. G. et al. The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*. **eLife**, v. 4, p. e08347–e08347, 2015.

LATIN AMERICAN GENOMIC SURVEILLANCE ARBOVIRAL NETWORK et al. Field and classroom initiatives for portable sequence-based monitoring of dengue virus in Brazil. **Nature Communications**, v. 12, n. 1, p. 2296, abr. 2021.

LI, P. An Empirical Evaluation of Four Algorithms for Multi-Class Classification: Mart, ABC-Mart, Robust LogitBoost, and ABC-LogitBoost. **arXiv:1001.1020 [cs]**, 7 jan. 2010.

LIPOVETSKY, S.; CONKLIN, M. Analysis of regression in game theory approach. **Applied Stochastic Models in Business and Industry**, v. 17, n. 4, p. 319–330, out. 2001.

LOPES, N.; NOZAWA, C.; LINHARES, R. E. C. Características gerais e epidemiologia dos arbovírus emergentes no Brasil. **Revista Pan-Amazônica de Saúde**, v. 5, n. 3, p. 10–10, 2014.

LUBKE, G. et al. Gradient Boosting as a SNP Filter: an Evaluation Using Simulated and Hair Morphology Data. **Journal of data mining in genomics & proteomics**, v. 4, 20 out. 2013.

LUNDBERG, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. **Nature Biomedical Engineering**, v. 2, n. 10, p. 749–760, out. 2018.

LUNDBERG, S. M.; ERION, G. G.; LEE, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. **arXiv:1802.03888 [cs, stat]**, 11 fev. 2018.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. Em: GUYON, I. et al. (Eds.). **Advances in neural information processing systems 30**. [s.l.] Curran Associates, Inc., 2017. p. 4765–4774.

LUNDBERG, S. M.; LEE, S.-I. Consistent feature attribution for tree ensembles. **arXiv:1706.06060 [cs, stat]**, 16 fev. 2018.

MARES-GUIA, M. A. M. DE M. et al. Yellow fever epizootics in non-human primates, Southeast and Northeast Brazil (2017 and 2018). **Parasites & Vectors**, v. 13, n. 1, 2020.

MARTINEZ, D. R. et al. Antigenic variation of the dengue virus 2 genotypes impacts the neutralization activity of human antibodies in vaccinees. **Cell reports**, v. 33, n. 1, p. 108226, 2020.

MINISTÉRIO DA SAÚDE DO BRASIL. **Vigilância em saúde no Brasil 2003|2019: da criação da Secretaria de Vigilância em Saúde aos dias atuais**. , 2019. Disponível em: <<https://www.saude.gov.br/images/pdf/2020/janeiro/20/Boletim-epidemiologico-SVS-02-1-.pdf>>

MINISTÉRIO DA SAÚDE DO BRASIL. **Monitoramento dos casos de arboviroses urbanas transmitidas pelo Aedes (dengue, chikungunya e Zika), Semanas Epidemiológicas 01 a 52. 2020**. , 2020. Disponível em: <<https://www.saude.gov.br/images/pdf/2020/janeiro/20/Boletim-epidemiologico-SVS-02-1-.pdf>>

MONATH, T. P. Yellow fever: an update. **The Lancet. Infectious Diseases**, v. 1, n. 1, p. 11–20, ago. 2001.

MONATH, T. P. Treatment of yellow fever. **Antiviral Research**, v. 78, n. 1, p. 116–124, abr. 2008.

Monitoramento dos casos de arboviroses urbanas transmitidas pelo Aedes (dengue, chikungunya e Zika), Semanas Epidemiológicas 01 a 52. 2020. Brasil. Ministério da Saúde., , [s.d.]. Disponível em: <<https://www.saude.gov.br/images/pdf/2020/janeiro/20/Boletim-epidemiologico-SVS-02-1-.pdf>>. Acesso em: 20 maio. 2020

MOORE, J. H.; ASSELBERGS, F. W.; WILLIAMS, S. M. Bioinformatics challenges for genome-wide association studies. **Bioinformatics**, v. 26, n. 4, p. 445–455, 15 fev. 2010.

MORAIS-RODRIGUES, F. **Microarray-based breast cancer classification using logistic regression and beyond**. [s.l.] Universidade Federal de Minas Gerais, 2017.

MORAIS-RODRIGUES, F. et al. Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression. **Gene**, v. 726, p. 144168, fev. 2020.

MORELI, M. L.; COSTA, V. A systematic review of molecular diagnostic methods for the detection of arboviruses in clinical specimens in Brazil and the importance of a differential diagnosis. **Virology Discovery**, v. 1, n. 1, p. 1, 2013.

MUKTAR, Y.; TAMERAT, N.; SHEWAFERA, A. Aedes aegypti as a Vector of Flavivirus. **J Trop Dis**, v. 4, n. 223, p. 2, 2016.

MUSSO, D.; GUBLER, D. J. Zika virus. **Clinical microbiology reviews**, v. 29, n. 3, p. 487–524, 2016.

PAHO/WHO. **Laboratory Diagnosis of Yellow Fever Virus infection - 2018 - PAHO/WHO | Pan American Health Organization**. Disponível em: <<https://www.paho.org/en/documents/laboratory-diagnosis-yellow-fever-virus-infection-2018>>. Acesso em: 23 jan. 2022.

PAHO/WHO. **Dengue - PAHO/WHO | Pan American Health Organization**. Disponível em: <<https://www.paho.org/en/topics/dengue>>. Acesso em: 24 jan. 2022a.

PAHO/WHO. **Epidemiological Update: Yellow Fever - 28 December 2021 - PAHO/WHO | Pan American Health Organization**. Disponível em: <<https://www.paho.org/en/documents/epidemiological-update-yellow-fever-28-december-2021>>. Acesso em: 23 jan. 2022b.

PAHO/WHO. **Yellow fever**. Disponível em: <<https://www.paho.org/en/topics/yellow-fever>>. Acesso em: 23 jan. 2022.

PAULES, C. I.; FAUCI, A. S. Yellow Fever — Once Again on the Radar Screen in the Americas. **New England Journal of Medicine**, v. 376, n. 15, p. 1397–1399, 13 abr. 2017.

PIERSON, T. C.; DIAMOND, M. S. The continued threat of emerging flaviviruses. **Nature microbiology**, v. 5, n. 6, p. 796–812, 2020.

QI, Y. Random forest for bioinformatics. Em: **Ensemble machine learning**. [s.l.] Springer, 2012. p. 307–323.

RIBEIRO, M. T.; SINGH, S.; GUESTIN, C. “**Why Should I Trust You?**”: **Explaining the Predictions of Any Classifier**. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. **Anais...** Em: THE 22ND ACM SIGKDD

INTERNATIONAL CONFERENCE. San Francisco, California, USA: ACM Press, 2016. Disponível em: <<http://dl.acm.org/citation.cfm?doid=2939672.2939778>>. Acesso em: 23 jul. 2019

RICE, C. M. et al. Nucleotide Sequence of Yellow Fever Virus: Implications for Flavivirus Gene Expression and Evolution. **Science (American Association for the Advancement of Science)**, v. 229, n. 4715, p. 726–733, 1985.

SAN MARTÍN, J. L. et al. The epidemiology of dengue in the Americas over the last three decades: a worrisome reality. **The American journal of tropical medicine and hygiene**, v. 82, n. 1, p. 128, 2010.

SCHAPIRE, R. E. The Boosting Approach to Machine Learning: An Overview. Em: DENISON, D. D. et al. (Eds.). **Nonlinear Estimation and Classification**. Lecture Notes in Statistics. New York, NY: Springer New York, 2003. v. 171p. 149–171.

SHEARER, F. M. et al. Global yellow fever vaccination coverage from 1970 to 2016: an adjusted retrospective analysis. **The Lancet infectious diseases**, v. 17, n. 11, p. 1209–1217, 2017.

SHRIKUMAR, A.; GREENSIDE, P.; KUNDAJE, A. Learning Important Features Through Propagating Activation Differences. **arXiv:1704.02685 [cs]**, 9 abr. 2017.

SILVA, N. I. O. et al. Recent sylvatic yellow fever virus transmission in Brazil: the news from an old disease. **Virology Journal**, v. 17, n. 1, p. 9, 23 jan. 2020.

SIMMONDS, P. et al. ICTV virus taxonomy profile: Flaviviridae. **The Journal of general virology**, v. 98, n. 1, p. 2, 2017.

ŠTRUMBELJ, E.; KONONENKO, I. Explaining prediction models and individual predictions with feature contributions. **Knowledge and Information Systems**, v. 41, n. 3, p. 647–665, dez. 2014.

VASCONCELOS, P. F. DA C. Yellow fever in Brazil: thoughts and hypotheses on the emergence in previously free areas. **Revista de Saúde Pública**, v. 44, n. 6, p. 1144–1149, dez. 2010.

VASCONCELOS, P. F.; MONATH, T. P. Yellow fever remains a potential threat to public health. **Vector-Borne and zoonotic diseases**, v. 16, n. 8, p. 566–567, 2016.

WANG, W. Y. S. et al. Genome-wide association studies: theoretical and practical concerns. **Nature Reviews Genetics**, v. 6, n. 2, p. 109, fev. 2005.

WHO. **Yellow fever**. Disponível em: <<https://www.who.int/news-room/fact-sheets/detail/yellow-fever>>. Acesso em: 23 jan. 2022.

WHO SCIENTIFIC GROUP ON ARBOVIRUSES. **Arboviruses and human disease : report of a WHO scientific group [meeting held in Geneva from 26 September to 1 October 1966].** : World Health Organization technical report series ; no. 369. World Health Organization, , 1967.

WOLPERT, D. H. The Lack of A Priori Distinctions Between Learning Algorithms. **Neural Computation**, v. 8, n. 7, p. 1341–1390, out. 1996.

WU, P. et al. Arbovirus lifecycle in mosquito: acquisition, propagation and transmission. **Expert reviews in molecular medicine**, v. 21, 2019.

Anexo

Artigos publicados

SALGADO, Á. et al. Machine learning models exploring characteristic single-nucleotide signatures in yellow fever virus. **PLOS ONE**, v. 17, n. 12, p. e0278982, 12 dez. 2022.

ADELINO, T. É. R. et al. Field and classroom initiatives for portable sequence-based monitoring of dengue virus in Brazil. **Nature communications**, v. 12, n. 1, p. 1–12, 2021.

GIOVANETTI, M. et al. Promoting Responsible Research and Innovation (RRI) during Brazilian activities of genomic and epidemiological surveillance of arboviruses. **Frontiers in Public Health**, v. 9, p. 894, 2021.

GOES DE JESUS, J. et al. Persistence of chikungunya ECSA genotype and local outbreak in an upper medium class neighborhood in Northeast Brazil. **PloS one**, v. 15, n. 1, p. e0226098, 2020.

LOPES, E. N. et al. Betacoronaviruses genome analysis reveals evolution toward specific codons usage: Implications for SARS-CoV-2 mitigation strategies. **Journal of medical virology**, v. 93, n. 9, p. 5630–5634, 2021.

NAVARRO BORBA, M. M. et al. In silico analysis of human T-lymphotropic virus type 1 complete genomes from patients with different clinical outcomes. **Future Virology**, v. 17, n. 3, p. 149–157, 2022.

NUNES, E. DA C. et al. Untargeted Metabolomics Insights into Newborns with Congenital Zika Infection. **Pathogens**, v. 10, n. 4, p. 468, 2021.

TOSTA, S. F. DE O. et al. Multi-epitope based vaccine against yellow fever virus applying immunoinformatics approaches. **Journal of Biomolecular Structure and Dynamics**, v. 39, n. 1, p. 219–235, 2021.

XAVIER, J. et al. The ongoing COVID-19 epidemic in Minas Gerais, Brazil: insights from epidemiological data and SARS-CoV-2 whole genome sequencing. **Emerging microbes & infections**, v. 9, n. 1, p. 1824–1834, 2020.

Capítulo de livro

GIOVANETTI, M. et al. Pan-genomics of virus and its applications. Em: **Pan-genomics: Applications, Challenges, and Future Prospects**. [s.l.] Elsevier, 2020. p. 237–250.