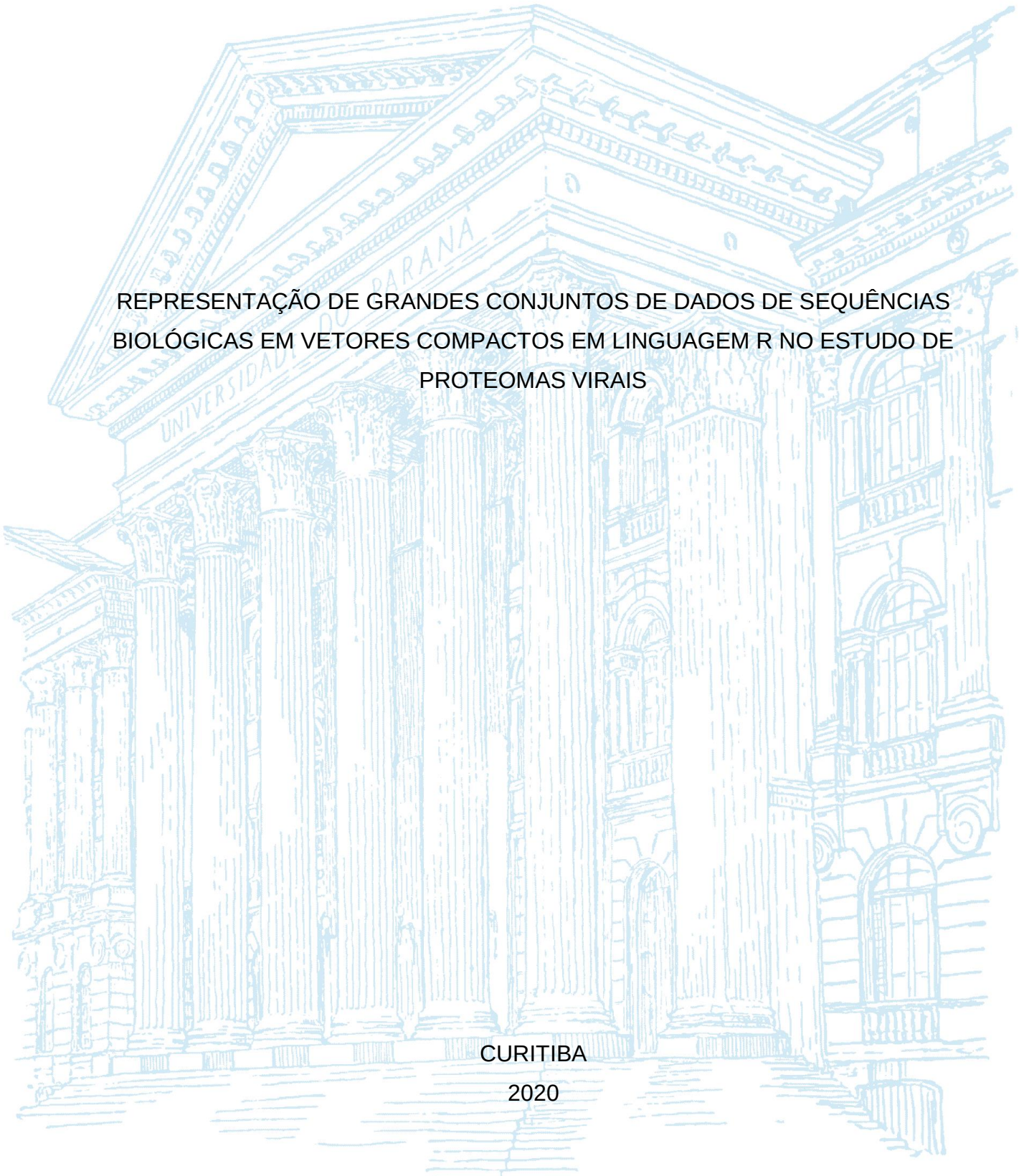
UNIVERSIDADE FEDERAL DO PARANÁ

DANRLEY RAFAEL FERNANDES



REPRESENTAÇÃO DE GRANDES CONJUNTOS DE DADOS DE SEQUÊNCIAS BIOLÓGICAS EM VETORES COMPACTOS EM LINGUAGEM R NO ESTUDO DE PROTEOMAS VIRAIS

CURITIBA

2020

DANRLEY RAFAEL FERNANDES

REPRESENTAÇÃO DE GRANDES CONJUNTOS DE DADOS DE SEQUÊNCIAS BIOLÓGICAS EM VETORES COMPACTOS EM LINGUAGEM R NO ESTUDO DE PROTEOMAS VIRAIS

Monografia apresentada ao curso de Graduação em Ciências Biológicas, Setor de Ciências Biológicas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Bacharel em Ciências Biológicas.

Orientador: Prof.° Dr.° Roberto Tadeu Raittz

Coorientadora: M.ª Camilla Reginatto De Pierri

CURITIBA

2020

TERMO DE APROVAÇÃO

DANRLEY RAFAEL FERNANDES

REPRESENTAÇÃO DE GRANDES CONJUNTOS DE DADOS DE SEQUÊNCIAS BIOLÓGICAS EM VETORES COMPACTOS EM LINGUAGEM R NO ESTUDO DE PROTEOMAS VIRAIS

Monografia apresentada ao curso de Graduação em Ciências Biológicas, Setor de Ciências Biológicas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Bacharel em Ciências Biológicas

_____

Orientador: Prof.° Dr.° Roberto Tadeu Raittz

Setor de Educação Profissional e Tecnológica, SEPT, UFPR

_____

Profa. Dra. Ana Claudia Bonatto

Departamento de Genética, UFPR

_____

Msc. Camila Pereira Perico

Programa de Pós-graduação Associado em Bioinformática, UFPR

Curitiba, 18 de Dezembro de 2020.

À Bruna Moura, pessoa com quem amo partilhar a vida. Com você tenho me sentido vivo de verdade. Obrigado pelo carinho, a paciência e por sua capacidade de me trazer paz na correria de cada semestre.

AGRADECIMENTOS

Inicialmente, gostaria de agradecer a minha família e amigos. Especialmente, a minha mãe e padrasto que sempre me apoiaram com tudo que eu precisava durante a minha vida, e minha namorada Bruna Moura por me ouvir em momentos difíceis.

Ao Prof.Dr. Roberto Tadeu Raittz eu agradeço a orientação incansável e a confiança que tornaram possível a realização desse projeto. A minha coorientadora Msc. Camila Reginatto De Pierri pelo apoio durante todo o processo de construção desse trabalho.

Ao Programa De Pós-Graduação Em Bioinformatica, juntamente, com a UFPR e demais instituições de ensino, que me proporcionaram a oportunidade de possuir um ensino superior e a expansão de meus horizontes. Ao Laboratorio de Inteligencia Artificial em Bioinformatica, por me apresentar oportunidades unicas e expandir minha visão de mundo.

A todos os amigos que direta ou indiretamente participaram da minha formação, o meu muito eterno agradecimento. Obrigada! pela contribuição valiosa durante essa jornada.

“By giving our students practice in talking with others, we give them frames for thinking on their own.” (Lev Vygotsky, 1978)

RESUMO

O pacote rSWeeP é uma implementação R do modelo SWeeP, projetado para lidar com BigData. O rSweeP atende à crescente demanda por métodos eficientes de representação heurística no campo da Bioinformática, em plataformas acessíveis a toda comunidade científica. Exploramos a implementação de rSWeeP usando um conjunto de dados contendo 31.386 proteomas virais, realizando análises filogenéticas e de componentes principais. Como estudo de caso, analisamos as cepas virais mais próximas do SARS-CoV, responsáveis pela recente pandemia de COVID-19, demonstrando que o rSWeeP pode classificar com precisão os organismos taxonomicamente. O pacote rSWeeP está disponível gratuitamente em https://bioconductor.org/packages/ release / bioc / html / rSWeeP.html.

Palavras-chave: Análise de sequências biológicas. Filogenia. Aminoácidos. Técnicas vetoriais. Técnicas livres de alinhamento.

ABSTRACT

The rSWeeP package is an R implementation of the SWeeP model, designed to handle BigData. rSweeP meets to the growing demand for efficient methods of heuristic representation in the field of Bioinformatics, on platforms accessible to the entire scientific community. We explored the implementation of rSWeeP using a dataset containing 31,386 viral proteomes, performing phylogenetic and principal component analysis. As a case study we analyze the viral strains closest to the SARS-CoV, responsible for thecurrent pandemic of COVID-19, confirming that rSWeeP can accurately classify organisms taxonomically. rSWeeP package is freely available at https://bioconductor.org/packages/ release/bioc/html/rSWeeP.html.

Keywords: Analysis of biological sequences. Phylogeny. Amino acids. Vectorial techniques. Free alignment techniques.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS OU SIGLAS

SWeeP	- Spaced Words Projection
rSWeeP	- Spaced Words Projection em R
NCBI	- Nacional Center of Biological Information
ICTV	- Committee on Taxonomy of Viruses
ssRNA	- Single strain R*ibonucleic Acid*
ssDNA	- Single strain *Deoxyribonucleic Acid*
dsDNA	- Double strain *Deoxyribonucleic Acid*
SARS	- *Severe Acute Respiratory Syndrome*
DNA	- Ácido desoxirribonucleico
RNA	- Ácido ribonucleico
UPGMA	- *Unweighted Pair Group Method with Arithmetic Mean*
NJ	- Neighbor Joining
COVID-19	- Corona Virus Disease of 2019
MERS	- MIddle East Respiratory Syndrome
CRAN	- Comprehensive R Archive Network
SARS-COV	- Severe Acute Respiratory Syndrome Coronavirus
PCA	- Principal Component Analysis

LISTA DE SÍMBOLOS

@ - arroba

® - Marca registrada

SUMÁRIO

1 INTRODUÇÃO

Análises de Bioinformática em grandes conjuntos de dados de sequências biológicas são comuns na atualidade, porém, dependendo do método a ser aplicado nas análises, leva um tempo considerável para a obtenção de resultados (DE PIERRI *et al.*, 2020). Na última década, as técnicas vetoriais e livres de alinhamento tiveram grande destaque na comparação e representação de sequências biológicas por sua eficiência e agilidade quando comparadas a maioria dos métodos heurísticos baseados em alinhamento (ZIELEZINSKI *et al.*, 2017).

Ferramentas livres de alinhamento são bem-sucedidas na análise de grandes conjuntos de dados (DE PIERRI *et al*., 2020; ZHANG *et al.*; 2017; Li *et al.*; 2017) entretanto, por serem técnicas relativamente recentes, parte da comunidade científica ainda é resistente quanto ao uso de tais técnicas. Mesmo com a eficácia deste já comprovada, as técnicas dependentes de alinhamento ainda são prioritárias nas análises de sequências biológicas.

Ferramentas que têm como base o mapeamento de palavras (*k-mers*) em espaços vetoriais têm sido o assunto de vários estudos recentes (DE PIERRI *et al*., 2020; ZHANG *et al.*; 2017; Li *et al.*; 2017; LEIMEISTER; 2014; HORWEGE; 2014; NOÉ; 2014; VINGA; 2014). A representação de sequências biológicas em espaços vetoriais possibilita a manipulação de grandes conjuntos de dados com agilidade, além de facilitar a aplicação de técnicas de mineração de dados (VINGA, 2014).

Spaced Words Projection (*SWeeP*) é um modelo computacional que permite a representação de informações de sequencias biológicas por meio de projeção vetorial. Recentemente, o modelo *SWeeP* teve sua eficiência demonstrada na representação de grandes conjuntos de dados e na construção de árvores filogenéticas em dois estudos em publicação recente. Um destes estudos envolveu proteomas mitocondriais e bacterianos (DE PIERRI *et al*., 2020), e o outro, referente ao presente trabalho, proteomas virais (FERNANDES *et al*., 2020).

Neste trabalho é apresentado o r*SWeeP*, uma implementação do modelo *SWeeP* em linguagem de programação R/Bioconductor. O modelo *SWeeP* é originalmente implementado em MATLAB®, que é uma ferramenta paga. A inclusão do *SWeeP* em outras plataformas, como é o caso do R/Bioconductor (software

gratuito e amplamente utilizado para a análise de dados genômicos) oportuniza o acesso de toda a comunidade científica à ferramenta.

Para testar a efetividade da implementação r*SWeeP*, foi realizado teste de performance, construção de árvores filogenéticas e Análise de Componentes Principais (PCA). Foi utilizado como estudo de caso 31.386 proteomas virais retirados do banco de dados *Nacional Center of Biological Information* (NCBI).

Os resultados mostraram que o r*SWeeP* manteve a efetividade já apresentada pelo modelo *SWeeP* original na análise de grandes conjuntos de dados biológicos. Em relação à inferência filogenética do grupo de estudo, r*SWeeP* gerou a maior árvore filogenética de proteomas virais encontradas na literatura atualmente. Foi observando o padrão de agrupamento conforme o tipo organizacional do ácido nucleico: Ácido Ribonucleico fita simples (ssRNA), Ácido Desoxirribonucleico fita simples (ssDNA) e Ácido Desoxirribonucleico fita dupla (dsDNA). Estes agrupamentos foram posteriormente reforçados pela análise de PCA. Foi identificado ainda neste estudo a proximidade entre o vírus causador do COVID19, responsável pela atual pandemia de Síndrome Respiratória Aguda Grave (SARS), e o vírus responsável pela pandemia de SARS em 2003, sugerindo que são o mesmo coronavírus, conforme reportado em outro estudo (GORBALENYA; 2020).

1.1 OBJETIVOS

1.1.1 Objetivo geral

- Disponibilizar uma implementação gratuita do modelo *SWeeP*, utilizando a plataforma R e o repositório Bioconductor.

1.1.2 Objetivos específicos

- Comprovar a viabilidade do pacote de funções r*SWeeP* como plataforma do método *SWeeP*
- Possibilitar o acesso de uma maior parcela da comunidade científica ao método *SWeeP*.

- Realizar análises filogenéticas utilizando os vetores gerados pelo *SWeeP* em linguagem de programação R.
- Aplicar técnicas de mineração de dados a resultados gerados pelo método *SWeeP*.
- Auxiliar na consolidação de técnicas livres de alinhamento como ferramentas para análises de sequências biológicas.
- Gerar uma árvore filogenética em linguagem de programação R com o método *SWeeP* a partir de um grande conjunto de dados.
- Trazer elementos que auxiliem no entendimento da filogenia global do grupo dos vírus.
- Observar a proximidade genética entre o coronavírus responsável pelo atual surto mundial de SARS-COV e outras estirpes virais

2 REVISÃO DE LITERATURA

2.1 VÍRUS

Vírus são os menores e mais simples organismos registrados pelo homem, tendo diâmetros que variam entre 15 e 22 nanômetros (nm) na menor família (*Circoviridae*) e entre 200 e 450 nm na maior família (*Poxviridae*) (KNIOE; 2007). São importantes agentes patogênicos em humanos, podendo causar doenças leves, letais e desencadear o desenvolvimento de diversos tipos de câncer(CHAPMAN; REISS; 1992). O estudo deste táxon é fundamental, visto que também está associado a fatores econômicos, pois são responsáveis por diversas doenças em animais e cultivares (CHAPMAN; REISS; 1992).

Os vírus não possuem metabolismo próprio e necessitam das células de outros organismos para se multiplicarem, portanto, são parasitas intracelulares obrigatórios (KNIOE; 2007). Fora de uma célula viva os vírus são apenas estruturas químicas (KNIOE; 2007). Este táxon não possui organelas, apenas genoma viral, portanto não são considerados células. A união do material genético mais o capsídio, e eventuais componentes proteicos, formam a partícula vírica (ou virion), que tem como função proteger e projetar em células hospedeiras o genoma do vírus (CHAPMAN; REISS; 1992). Embora a maioria destes organismos possa ser vista apenas por microscopia eletrônica, existem exceções, a exemplo a família *Poxviridae*, cujos gêneros podem ser vistos por microscopia ótica (KNIOE; 2007). Os vírus podem variar em tamanho, composição e estrutura, entretanto, todos materiais genéticos virais possuem uma funcionalidade em comum: alguma tática de síntese da RNA polimerase. Esta enzima é necessária para o processo de sequestro da maquinaria celular de outros organismos (CHAPMAN; REISS; 1992).

De acordo com a estrutura do virion, os vírus se dividem em dois grupos principais: sem envelope viral ou com envelope viral (FIGURA **Erro! Fonte de referência não encontrada.**). Os vírus sem envelope possuem apenas o capsídeo, composto por proteínas simples, recobrindo curtos fragmentos genéticos. Já os vírus envelopados geralmente possuem genomas longos, associados com a transcrição de várias proteínas, um capsídeo com composição proteica mais complexa e um envoltório lipoproteico, denominado envelope viral (KNIOE; 2007).

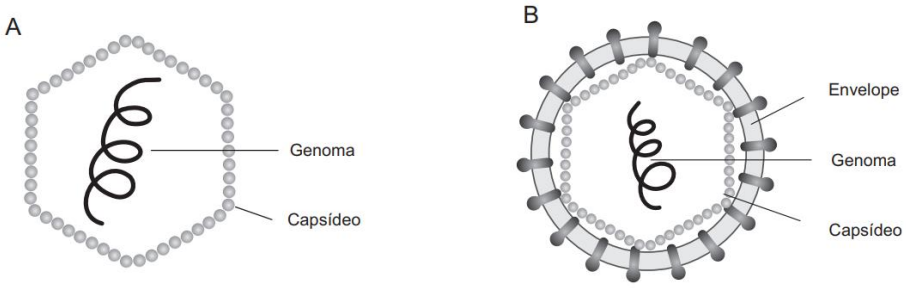


FIGURA 1 – ESTRUTURA FUNDAMENTAL DAS PARTÍCULAS VÍRICAS E SEUS COMPONENTES.
FONTE: Adaptado de KNIOE; HOWLEY (2007).
LEGENDA: Representação esquemática de um virion sem envelope (A) e com envelope (B).

2.1.1 Genoma viral

O genoma viral pode ser constituído tanto por ácido ribonucleico (RNA) quanto por desoxirribonucleico (DNA), mas nunca pelos dois. Por isso, vírus são comumente denominados RNA vírus ou DNA vírus. A maioria das famílias virais contêm apenas uma cópia do genoma por virion (são haploides), porem os retrovírus, possuem duas cópias do genoma (são diploides). Além disso, a organização genômica, tamanho, e número de genes são muito variados no táxon (KNIOE; 2007).

Os vírus possuem organizações gênicas únicas e seu material genético pode ser mantido na forma de múltiplas moléculas (KNIOE; 2007). O genoma viral pode apresentar variações: DNA de fita simples (ssDNA), DNA de fita dupla (dsDNA), RNA de fita dupla (dsRNA) e RNA de fita simples (ssRNA). Além disso um virion pode manter seu material genético na forma de múltiplas moléculas (KNIOE; 2007).

Frequentemente o genoma viral está associado proteínas covalentemente e negativamente carregadas, histônicas por exemplo, facilitando a ligação em ácidos nucleicos positivamente carregados (CHAPMAN; REISS; 1992).

2.1.2 Taxonomia viral

Durante o congresso internacional de microbiologia de 1996, em Moscou, foi criado o Comitê Internacional de Taxonomia Viral (do inglês, International Committee on Taxonomy of Viruses - ICTV) com o objetivo de organizar um esquema único e

universal para classificar taxonomicamente os vírus (SIDDELL; DAVISON *et al.* 2019).

Devido a tarefa de regulação do código de nomenclatura e de avaliação de novos táxons ser de responsabilidade quase exclusiva do ICTV, a taxonomia viral tem características singulares. Os nomes das espécies deste grupo normalmente derivam do nome popular dado em inglês e a taxonomia utiliza critérios politéticos para definir os indivíduos (LEFKOWITZ *et al*., 2018). Além disso, o táxon viral não é monofilético, não existe nenhum conjunto fixo de características que possa ser atribuído a todos os vírus. Para admissão de um novo membro ao táxon, é feita a comparação do candidato com um vírus já consolidado e de características similares (KARUPIAH; 2002). Embora seja necessário a avaliação de muitas características, elas geralmente se resumem à atributos morfológicos (KARUPIAH; 2002).

O nível taxonômico, acima de espécie, é definido de maneira relativa pelo ICTV: Um gênero só é formado quando existe um conjunto de espécies apresentando certas características comum, uma família ou subfamília só é formada quando existe um conjunto de espécies apresentando certas características comuns, e assim por diante (KARUPIAH; 2002). Nem todos os níveis taxonômicos precisam ser usados para um determinado agrupamento de vírus, portanto, enquanto a maioria das espécies é agrupada em gêneros e gêneros em famílias, nem todas as famílias contêm subfamílias e apenas algumas famílias são agrupadas em ordens (LEFKOWITZ *et al*., 2018) Consequentemente, a família é o grupo taxonômico mais alto usado de forma consistente e que apresenta a descrição mais generalizada de um determinado grupo de vírus, sendo a referência do sistema taxonômico (KARUPIAH; 2002).

Entretanto, o método de classificação empregado pelo ICTV tem sido considerado por diversos taxinomistas confuso e controverso (CALISHER; 2003). Para corrigir erros taxonômicos o ICTV recorre aos próprios membros do comitê. E esse processo já apresentou erros classificatórios (CALISHER; 2003). Visto que uma taxonomia organizada e confiável é do interesse de toda comunidade científica, ainda é preciso avançar neste campo, desenvolvendo e disponibilizando tecnologias para auxiliar na resolução dos problemas de classificação taxonômica.

2.1.3 Síndrome Respiratória Aguda Grave de Coronavírus (SARS-COV)

Em dezembro de 2019, foi relatado pela comissão Municipal de Saúde de Wuhan, na China, um surto de pneumonia viral causada por um patógeno desconhecido (World Health Organization, 2020). Posteriormente, o patógeno desconhecido foi identificado como um vírus da família coronavirus e denominado 2019-nCoV pela Organização Mundial da Saúde (OMS).

Este patógeno possui caracteristicamente um genoma composto por aproximadamente 30000 nucleotídeos, codificando quatro proteínas estruturais, além de algumas proteínas não estruturais (Figura 2):

a) Proteína do Nucleocapsídeo: Reguladora do processo de replicação viral (BOOPATHI; POMA; KOLANDAIVEL; 2020);
b) Proteína de membrana: Abundante na superfície viral. Acredita-se que sua função seja organizacional na montagem do coronavírus (KIRCHDOERFER *et al*.; 2016);
c) Proteína Spike: Está na superfície do vírus, e tem as funções de mediar a ligação do vírus com as células hospedeiras e é responsável pela fusão entre as membranas virais e hospedeira (KIRCHDOERFER *et al*.; 2016);
d) Proteína do envelope: É a menor estrutura da partícula viral, com cerca de 76 à 109 aminoácidos, ela é responsável pela montagem do vírus, sua proteção e favorece a interação vírus-célula hospedeira (GUPTA *et al*; 2020).

O mecanismo de entrada, replicação e empacotamento de RNA na célula hospedeira está mapeado na Figura *3*.

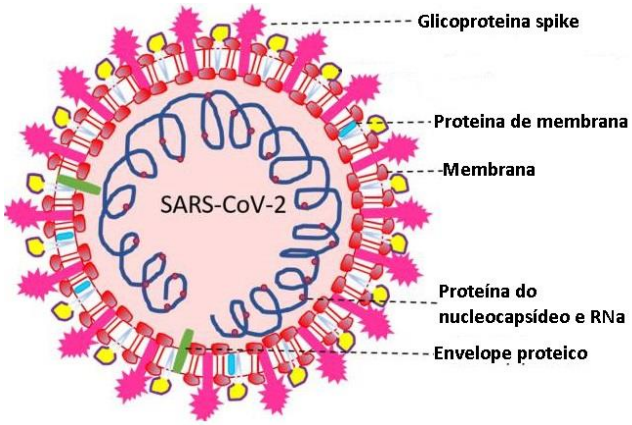


FIGURA 2 - *REPRESENTAÇÃO ESQUEMÁTICA DA ORGANIZAÇÃO DO GENOMA E DOMÍNIOS FUNCIONAIS DO SARS-COV2.*

FONTE: Adaptado de BOOPATHI; POMA; KOLANDAIVEL (2020).

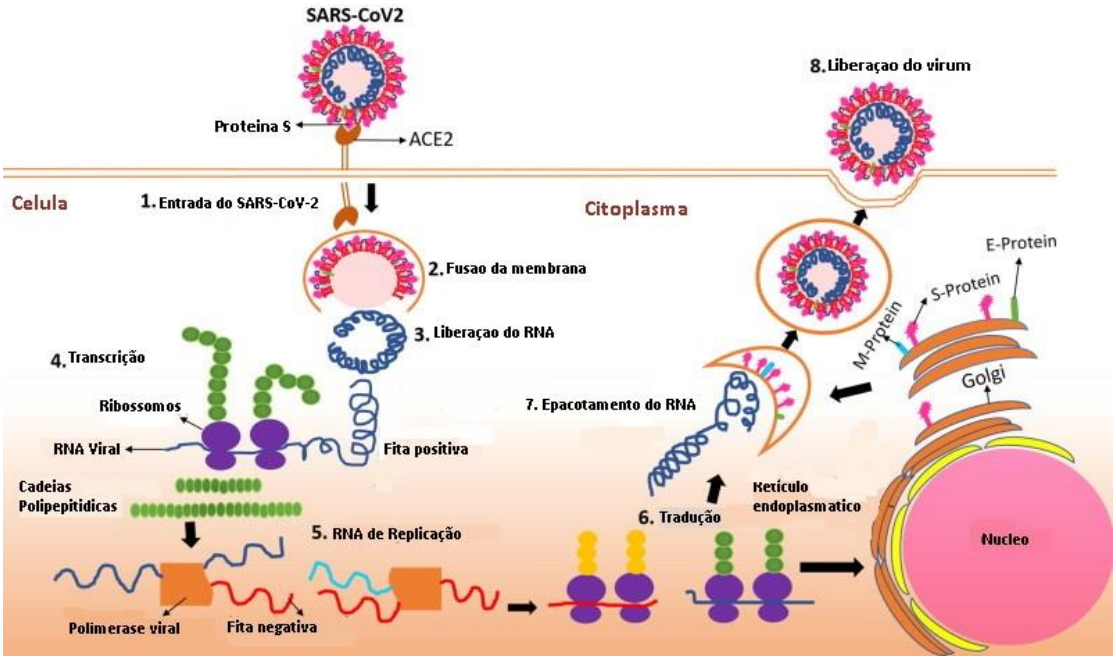


Figura 3 – O DIAGRAMA ESQUEMÁTICO DO MECANISMO DE ENTRADA DE COVID-19 E REPLICAÇÃO VÍRAL E EMPACOTAMENTO DE RNA VÍRAL NA CÉLULA HUMANA.

FONTE: Adaptado de BOOPATHI; POMA; KOLANDAIVEL (2020).

Os vírus do tipo coronavírus são caracterizados por terem RNA envelopados, com ampla variação fenotípica e genotípica (ZAK *et al*., 2012). Embora este grupo seja encontrado em diversas espécies animais, como pássaros, gatos, cães, porcos, ratos, cavalos, baleias e humanos, é mais comumente encontrado em morcegos (ZAK *et al*., 2012; CUI; LI; SHI, 2019). Nos diversos hospedeiros, este grupo viral é responsável por doenças respiratórias, entéricas, hepáticas e neurológica

(CUI; LI; SHI; 2019). Em humanos, seis estirpes de coronavírus causam doenças respiratórias, destas, quatro são conhecidos por serem endêmicos, porém pouco graves, e dois são responsáveis por surtos virais (ZAK *et al*., 2012). Até a ocorrência das epidemias de síndrome aguda respiratória (SARs) no ano de 2002 e 2003 (em Guangdong, China), em 2010 (no Oriente Médio) e a atual pandemia mundial de COVID-19 (do inglês, *Coronavirus disease* 2019), este grupo viral não era considerado altamente patogênico em humanos (CUI; LI; SHI; 2019).

De acordo com Cui Li e Shi (2019) é possível que o coronavírus responsável pela epidemia de síndrome respiratória que surgiu em países do Oriente Médio, denominada MERS-CoV (do inglês, *Middle East respiratory syndrome coronavirus*) originou-se em morcegos (hospedeiros naturais) e os camelos e dromedários foram os hospedeiros intermediários. Já os casos não endémico em humanos, originaram-se provavelmente de roedores. A cepa de coronavírus responsável pela a síndrome da diarreia aguda suína (SADS), teria como hospedeiro natural também o morcego (FIGURA 4).

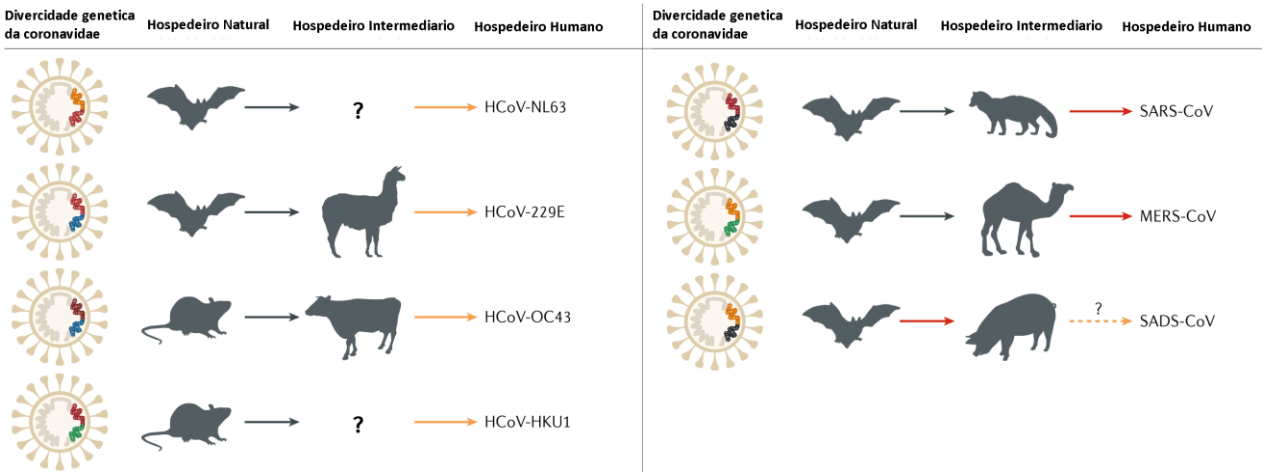


Figura 4 - ORIGENS ANIMAL DA CORONAVIRIDAE

FONTE: Adaptado de CUI; LI; SHI (2019).

LEGENDA: Setas quebradas indicam transmissão potencial entre espécies. Setas pretas indicam infecção em animais intermediários, setas amarelas indicam infeção leve em humanos e setas vermelhas indicam infecção grave em humanos ou animais.

2.2 ANÁLISES DE BIOINFORMÁTICA

Um dos objetivos do presente trabalho foi observar a proximidade genética entre o coronavírus responsável pelo atual surto mundial de SARS-COV e outras estirpes virais. Para tal, utilizamos ferramentas e métodos de Bioinformática.

A bioinformática multidisciplinar que combina elementos de diversos campos da ciência, como bioestatística, ciências biológicas, bioquímica, biologia molecular, genética, entre outras. Esta área busca facilitar análises de processos biológicos, interpretação dos dados e inferência biológica, interpretação dos dados e inferência biológica, uma possibilidade para estes estudos moleculares (MOORE, 2007). Neste trabalho foram realizadas análises de Bioinformática utilizando metodologias recentes para a comparação de sequencias biológicas, inferência filogenética, mineração de dados e análise de componentes principais (PCA).

2.2.1 Análise Comparativa de sequências genômicas

A comparação entre sequencias genômicas é a etapa inicial e fundamental dos estudos filogenéticos evolutivos (JUN *et al*., 2010). Para que uma comparação seja viável é preciso usar uma medida quantitativa de similaridade bem definida (SALEMI; VANDAMME, 2003). Os dois principais métodos para comparação de genomas: método de alinhamento (global ou local); e métodos livre de alinhamento (LEIMEISTER; MORGENSTERN, 2014). Em análises de grande escala, metodologias independentes de alinhamento são utilizadas para escapar das limitações de tempo e processamento dos métodos de alinhamento.

Um exemplo de método rápido de comparação de sequencias é a pesquisa por similaridade entre genes, “todos-contra-todos” (ARUNACHALAM *et al*., 2010; MAHMOOD *et al.*., 2012; VIALLE *et al*., 2016; NICHIO *et al*., 2019). Neste trabalho a comparação de sequencias é realizado por método livre de alinhamento, baseado na varredura de k-mers.

A metodologia de k-mer é empregada pela maioria dos estudos filogenéticos livres de alinhamento. O método de k-mer consiste no destaque da ocorrência de palavras de comprimento fixo “k” em um conjunto de sequencias biológicas (VINGA; ALMEIDA, 2003). A variação do comprimento das palavras geralmente varia de acordo com o tipo do conjunto de dados a ser analisado (SIMS *et al.*, 2009). A primeira etapa deste tipo de análise é o mapeamento e vetorização das sequencias

originais, por meio de uma janela deslizante, onde a resolução será determinada pelo comprimento da palavra (FIGURA **Erro! Fonte de referência não encontrada.**). Em seguida uma medida de distância deve ser definida para o intervalo entre vetores (VINGA; ALMEIDA, 2003). As medidas de distancia aplicadas podem variar, mas a mais utilizada é euclidiana, definida pela soma da raiz quadrada da diferença entre as coordenadas (BODEN *et al*., 2013).

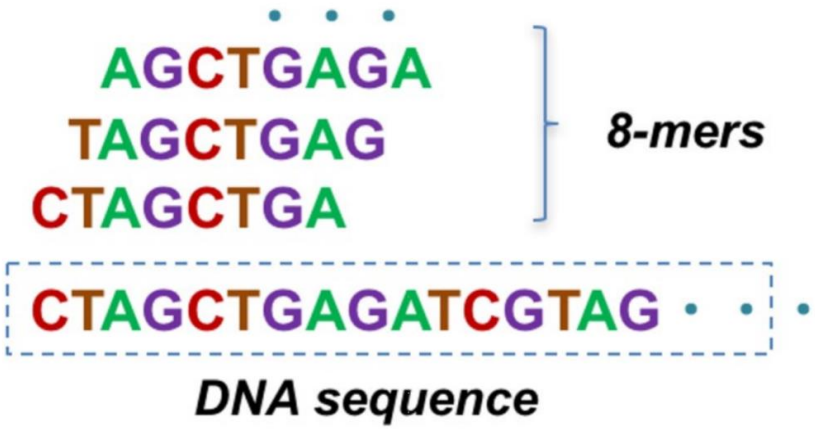


FIGURA 5 - VARREADURA DE SEQUÊNCIA COM JANELA DESLIZANTE DE TAMANHO K=8
FONTE: GE *et al*., (2020)
LEGENDA: Exemplo do processo de extração k-mer em uma dada uma sequência de DNA

Existem estudos onde ocorre a utilização do método de k-mers espaçados (Figura 6), mas baseados em um padrão fixo P que ignora posições. As palavras espaçadas são representadas em um vetor binário, a partir das frequências relativas, onde o cálculo da distância dos pares é aplicado (BODEN *et al.*, 2013; LEIMEISTER *et al*., 2014; HORWEGE *et al.*, 2014; DE PIERRI *et al*., 2020).

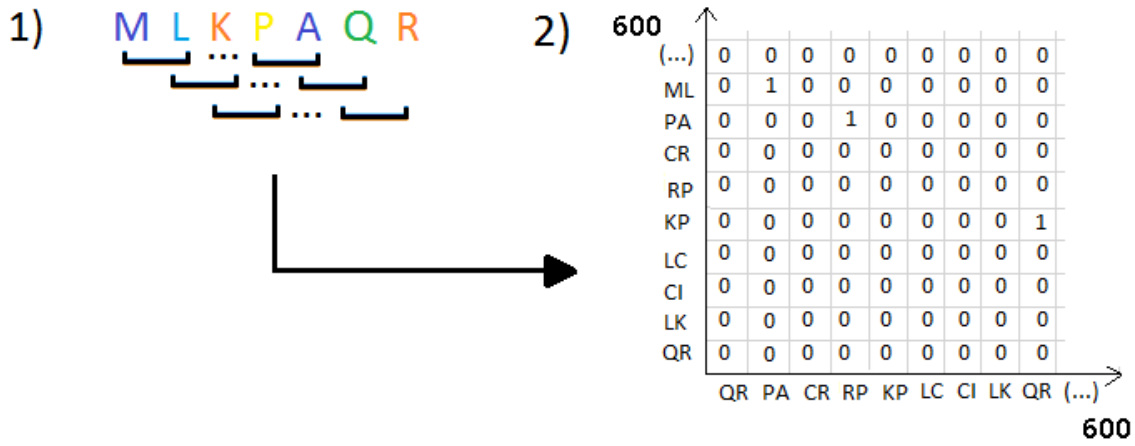


FIGURA 6 - Exemplo o método de k-mers espaçados

FONTE: PIERRI *et al*., (2017)

2.2.2 Análise filogenética

Antes do advento das tecnologias de sequenciamento de DNA, as árvores filogenéticas eram usadas quase que exclusivamente para representar taxonomias (SALEMI; VANDAMME, 2003). Com a consolidação de análises moleculares como ferramentas de reconstrução filogenética, as árvores de inferência filogenética têm aplicação em quase todos os ramos da biologia. Além de representar as relações entre as espécies, estas inferências são usadas na descrição de relações parálogas em famílias de genes, histórias populacionais, dinâmicas evolutivas e na epidemiologia de patógenos (YANG; RANNALA, 2012).

Os métodos filogenéticos revolucionaram muitos ramos da biologia, principalmente a genética populacional. Neste campo surgiu a genética computacional moderna, que permitiu a associação de dados com diferentes origens em uma única inferência (JUN *et al*., 2010). A utilização de genes únicos para análises evolutivas tornou-se obsoleto e passou a ser comum utilizar modelos com múltiplas sequências biológicas para realização de tais estudos (SAKAMOTO, 2016). Com isso, até mesmo para realização de associações entre características fenotípicas, passou a se levar em conta os traços filogenéticos, buscando assim evitar interpretações errôneas de contingências evolutivas (YANG; RANNALA, 2012).

Para inferência de árvores filogenéticas são utilizados modelos matemáticos. Assim como os dados escolhidos afetam a confiabilidade da filogenia, o modelo

matemático nela utilizado determina a precisão da mesma. De acordo com Delsuc, Brinkmann e Phillipe (2005), os três principais métodos matemáticos utilizados são:

a) Métodos baseados em parcimônia, que buscam formar a filogenia com o mínimo de mudanças possível (MADDISON *et al*., 1997);
b) Métodos baseados em probabilidades, que utilizam funções que calculam a probabilidade de uma determinada árvore ter produzido os dados observados, como a Máxima Verossimilhança e a Inferência Bayesiana (BRINKMANN; PHILLIPE, 2005)
c) Métodos baseados em distância, onde as distâncias sequenciais em pares são calculadas assumindo um modelo de substituição de nucleotídeos (FIGURA 2). Os principais métodos aplicados são o *Neighbor joining* (NJ) (SAITOU; NEI, 1978), o *Unweighted Pair Group Method with Arithmetic Mea*n (UPGMA) (MICHENER; SOKAL, 1957) e o Ward (WARD, 1963).

Neste trabalho foi utilizado o método Ward para inferir a filogenia do conjunto de dados virais. Este método baseia-se em clusterização hierárquica, hierarquizaçao baseada no cálculo da distância entre os objetos do conjunto de dados, onde a distâncias entre todas as sequências dos grupos são calculadas. Entretanto, existem diferentes maneiras de definir estas distâncias, resultando em tipos diferentes de clusters (HOLMES; HUBER, 2019).

2.2.3 Mineração de dados em espaço vetorial

A Mineração de dados é um processo que tem por objetivo reconhecer padrões e modelos em grandes conjuntos de dados. O desenvolvimento de metodologias de mineração de dados capazes de manipular informações biologias, é o principal meio para a bioinformática analisar grandes conjuntos de dados biológicos (ZAKI *et al*., 2007).

Métodos de aprendizado de máquina e de mineração de dados geralmente, quando aplicados ao reconhecimento de padrões em grandes volumes de dados biológicos, tem a possibilidade de reproduzir de maneira exata a complexidade

implícita dos sistemas biológicos (MOORE, 2007). Existem duas principais categorias para classificar métodos de reconhecimento de padrões: aprendizagem supervisionada, onde as classes são definidas, e aprendizagem não supervisionada, onde as classes são aprendidas de acordo com a similaridade dos padrões (RUSSELL; NORVIG, 2004).

A álgebra linear tem por objetivo estudar o comportamento de operações definidas sobre conjuntos, especificamente espaços vetoriais (PELLEGRINI, 2016). Um “vetor” é matematicamente definido como uma sequência ordenada de valores. Normalmente vetores são representados em segmentos de reta orientados em um espaço euclidiano de “n” dimensões, sendo a adição e multiplicação de vetores suas operações fundamentais (RUSSELL; NORVIG, 2004).

Umas das formas comuns para referenciar vetores em espaços virtuais é por escrita matricial, como uma matriz linha (com uma única linha) ou uma matriz coluna (com uma única coluna). Isto é possível pois operações vetoriais e operações matriciais possuem propriedades muito semelhantes (SANTOS, 2010). Um espaço vetorial permite a multiplicação e soma de elementos de forma escalar, assim, permite equações com elementos que não sejam do próprio espaço (PELLEGRINI, 2016).

2.2.4 Análise de Componentes principais

Dentro do campo das ciências biológicas muitas vezes os dados coletados de um experimento possuem dimensionalidades que são superiores ao numero de instâncias, como as medidas de expressão gênica e a espectrometria (RINGNÉR, 2008). Dados de alta dimensionalidade acabam tendo sua compreensão limitada por métodos analíticos simples (RINGNÉR, 2008). Neste sentido, a técnica Análise de Componentes Principais (PCA) é uma técnica para análise de problemas multivariados com grande aplicação multidisciplinar (PEARSON, 1901), além de ser muito provavelmente a primeira técnica criada para este tipo de problema (ABDI; WILLIAMS, 2010).

O PCA é um algoritmo de redução de dimensionalidade de dados, que permite a retenção da maior parte das informações dentro do conjunto (RINGNÉR, 2008). Para isso, a técnica identifica quais são os componentes com maior variância

entre os objetos. Este tipo de análise permite a criação de projeções com dimensionalidades selecionadas pelo analista, assim, facilitando identificar visualmente agrupamentos e dados de interesse (RINGNÉR, 2008).

3 MATERIAL E MÉTODOS

Neste trabalho foi utilizado a linguagem de programação R para explorar conceitos algébricos e implementação da ferramenta SWeeP (implementada originalmente em MATLAB®). Este trabalho visou realizar a implementação do modelo de projeção vetorial de sequências biológicas *SWeeP* em uma plataforma gratuita. O software desenvolvido, *rSWeeP*, foi explorado por meio da reconstrução filogenética e distribuição da taxonomia viral. Todos os processos foram executados em sistema operacional Windows 7, em computador *desktop* de processador Intel(R) Core(R) I5, 320 GHz e 12 GB de memória RAM.

3.1 O MODELO *SWEEP*

SWeeP é modelo computacional em bioinformática que realiza a representação de sequências biológicas através da projeção de vetores de pequena dimensionalidade (DE PIERRI *et al*., 2020). Para tal, o conceito de palavras espaçadas (BODEN *et al*., 2013; LEIMEISTER *et al.*, 2014) é utilizado para varrer sequências de amino ácidos, a fim de gerar vetores binários que representam as informações contidas nas sequências de origem. Estes vetores podem ter sua dimensionalidade alterada conforme o lema Johnson-Lindenstrauss (JOHNSON, LINDENSTRAUSS, 1986), sem que ocorra perda de informação. Neste trabalho, foi usado a máscara “11011” para a varredura das sequencias dos proteomas virais e a projeção matricial de tamanho 600x600.

3.2 CONJUNTO DE DADOS E INFERÊNCIA FILOGENÉTICA

Para explorar e testar a viabilidade do pacote r*SWeeP*, foram criados vetores *SWeeP* a partir de todas as sequências proteicas virais disponíveis no RefSeq NCBI, disponível em https://ftp.ncbi.nlm.nih.gov/genomes/Viruses/ (acesso em 6 de julho de 2019), em formato FASTA da proteína (.faa). As sequências foram concatenas a fim de gerar um único multiFASTA, representando o “proteoma viral” (processo realizado para cada organismo), conforme protocolo do estudo de DE PIERRI e colaboradores (2020).

Para a inferência filogenética estes vetores tiveram a distância Euclidiana entre eles calculada e foram agrupados pelo método Ward (WARD, 1963). As árvores filogenéticas encontram-se disponíveis em https://github.com/DanrleyRF/Suplementar.

4 RESULTADOS E DISCUSSÃO

Os resultados deste trabalho, apresentados nesse capitulo, foram publicados em forma de *Preprint* no BioRxiv (FERNANDES *et al*., 2020).

4.1 DESENVOLVIMENTO DE FERRAMENTA RSWEEP

A linguagem de programação R, por ser um software gratuito e amplamente utilizado para bioinformática, foi escolhida neste trabalho como a plataforma para implementação do modelo *SWeeP*. O pacote de funções r*SWeeP* foi criado com o intuito de conter todas as ferramentas necessárias para utilização do *SWeeP* em ambiente R. Este pacote possui duas funções principais:

a) OrthoBase: que gera uma matriz ortonormal de tamanhos específicos (definido pelo usuário).
b) SWeeP: que gera vetores *SWeeP* a partir de sequências proteicas.

O C*omprehensive R Archive Network* (CRAN) é o repositório de funções inato do Software R, contudo o Bioconductor é o depositário preferencial para os pacotes de funções voltados para bioinformática. Desta forma, o r*SWeeP* foi projetado segundo os critérios de padronização estipulados por este repositório, visando sua aceitação na plataforma.

Uma das principais exigências do repositório Bioconductor é compatibilidade entre os pacotes nele armazenados, assim o pacote r*SWeeP* é executado tanto com a entrada de um arquivo multiFASTA quanto de um objeto de classe “AAStringSet” contendo sequências de amino ácidos.

A função principal do pacote, *SWeeP*, para ser executada deve receber uma matriz ortonormal com 160 mil linhas, que pode ou não ser gerada pela função orthoBase (FIGURA **Erro! Fonte de referência não encontrada.**).

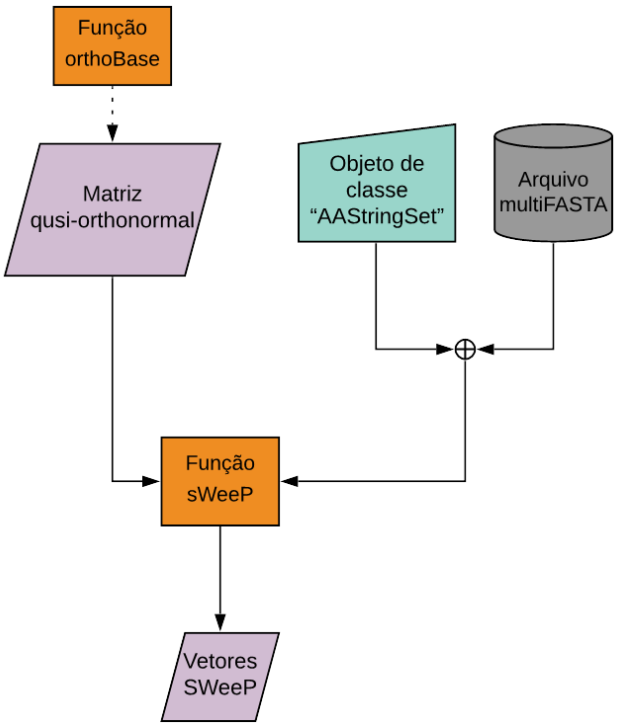


FIGURA 7 – FLUXOGRAMA DE FUNÇÕES DO PACOTE r*SWeeP*

FONTE: O autor (2020).

LEGENDA: Representação sistemática do funcionamento do pacote r*SWeeP*. A função *SWeeP*, recebe duas entradas, um arquivo multiFASTA ou objeto de classe “AAStringSet” com aminoácidos e uma matriz quasi-ortonormal com 160000 colunas e retorna uma matriz de vetores *SWeeP*.

A função principal do pacote, rSWeeP, recebe duas entradas, um arquivo de camada com proteínas e uma matriz quase ortonormal. Em seguida, cada sequência no multiFASTA é mapeada a partir dos aminoácidos contidos nas sequências, usando uma máscara "11011". Esses mapeamentos são projetados em uma base quase ortonormal, criando vetores SWeeP, que retêm todas as informações de comparação nas strings. Os vetores SWeeP podem ser usados para aprendizado de máquina em geral. A FIGURA **Erro! Fonte de referência não encontrada.** apresenta um tutorial de instalação do rSWeeP.

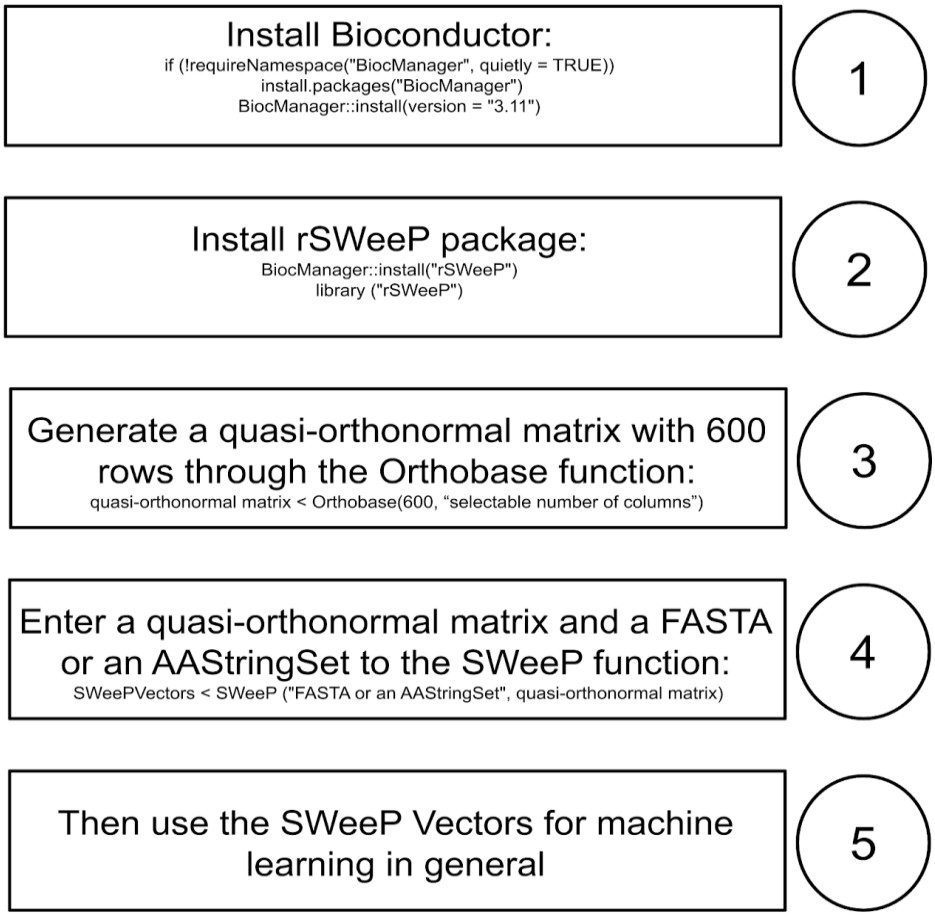


FIGURA 8 - Estratégias de utilização do pacote rSWeeP

FONTE: FERNANDES et al., 2020.

LEGENDA: A instalação do rSWeeP consiste em 5 etapas principais. Os processos padrões para construção dos vetores SWeeP são detalhados no estudo de DE PIERRI e colaboradores (2020).

O pacote rSWeeP está depositado na plataforma Bioconductor, portanto, é necessária a instalação do repositório Bioconductor. A execução do rSWeeP só é possível por meio das funções do Biocondutor. Para baixar a função, acesse https://github.com/DanrleyRF/Suplementar.

Foi obtido sucesso na disponibilização de um pacote R no Repositório Bioconductor, que permite a utilização do modelo *SWeeP* na análise de sequências biológicas (disponível em bioconductor.org/packages/release/bioc/html/rSWeeP.html) O pacote r*SWeeP* foi aceito pela plataforma Bioconductor no dia 23 de janeiro de 2020 e foi disponibilizado no repositório a partir da versão *Development 3.12.* Segundo os dados fornecidos pela própria plataforma, *d*esde a publicação do r*SWeeP* até o mês de novembro, última atualização dos dados, a ferramenta teve

763 downloads de 231 IP diferentes (TABELA **Erro! Fonte de referência não encontrada.**).

TABELA 1 – DOWLOADS DA FERRAMENTA r*SWeeP*

Mês	Número de IP distintos	Número de downloads
Jan/2020	1	1
Fev/2020	5	10
Mar/2020	2	8
Abr/2020	24	30
Mai/2020	43	81
Jun/2020	40	134
Jul/2020	29	75
Ago/202	45	114
Set/2020	36	103
Out/2020	54	117
Nov/2020	31	90
Total	231	763

FONTE: Bioconductor (2020).

4.2 CONSTRUÇÃO DE ÁRVORES FILOGENÉTICAS

Foram criados vetores *SWeeP* (com dimensionalidade tamanho 600) a partir de todos os proteomas virais disponíveis no NCBI, totalizando 31,386 (30k) diferentes estirpes, e uma super-árvore filogenética foi construída. Importante destacar que esta árvore é quase dez vezes maior que a referência encontrada na literatura (ZHANG; 2017).

A árvore construída apresenta um alto número de espécies virais. Enquanto existe espécies com um único represente, existem grupos, como o rotavírus A, com mais de mil estirpes. Assim, para uma melhor compreensão da distribuição taxonômica viral, usando os mesmos vetores, filtramos os dados, selecionando espécies de vírus classificado como "exemplos de ICTV (*International Committee on Viral Taxonomy*)" (LEFKOWITZ *et al*., 2018). A árvore filogenética filtrada com 4.833 proteomas virais (4k), foi construída utilizando o mesmo método e a mesma projeção vetorial da árvore total. Nesta árvore filtrada, estão representadas apenas uma amostra viral por espécie.

4.2.1 Teste de performance

A árvore de 30k foi gerada em duas horas e dezessete minutos, e a árvore 4k levou treze minutos. r*SWeeP* apresentou uma curva de crescimento linear para o tempo de formação do vetores SWeeP, como já era esperado, de quase 500 sequências comparadas por minuto. Desta forma, as funções apresentadas pelo pacote r*SWeeP* podem dar aos seus usuários poder computacional para administra grandes conjuntos de dados de maneira prática.

4.3 ANÁLISE FILOGENÉTICA GLOBAL

A fim de testar a eficácia dos dados gerados pelo pacote r*SWeeP*, nós projetamos graficamente os dois componentes principais dos vetores *SWeeP* gerados a partir dos 30k de proteomas virais, retirados do NCBI, e do conjunto de 4k de proteomas virais, classificadas como "*exemplares do ICTV*". Nesta plotagem observamos um padrão de agrupamento em relação ao tipo organizacional do ácido nucleico, é evidente a separação entre os três clados principais: ssRNA, ssDNA, dsDNA (classificados de acordo com o ICTV) Tanto para o conjunto 4k (FIGURA 8a) quanto para o grupo 30k (FIGURA 8b). Contudo, existem casos divergentes, principalmente dentro do agrupamento de vírus ssDNA, isso pode ter sido causado pela presença de grupos virais que não possuem clado bem definido no conjunto de dados. Casos de erro de anotação nos bancos de dados do NCBI podem ter favorecido estas discrepâncias, como também foi observado por Calisher e colegas (2006), sendo que vários dos erros por eles apontados continuam sem correção até o momento.

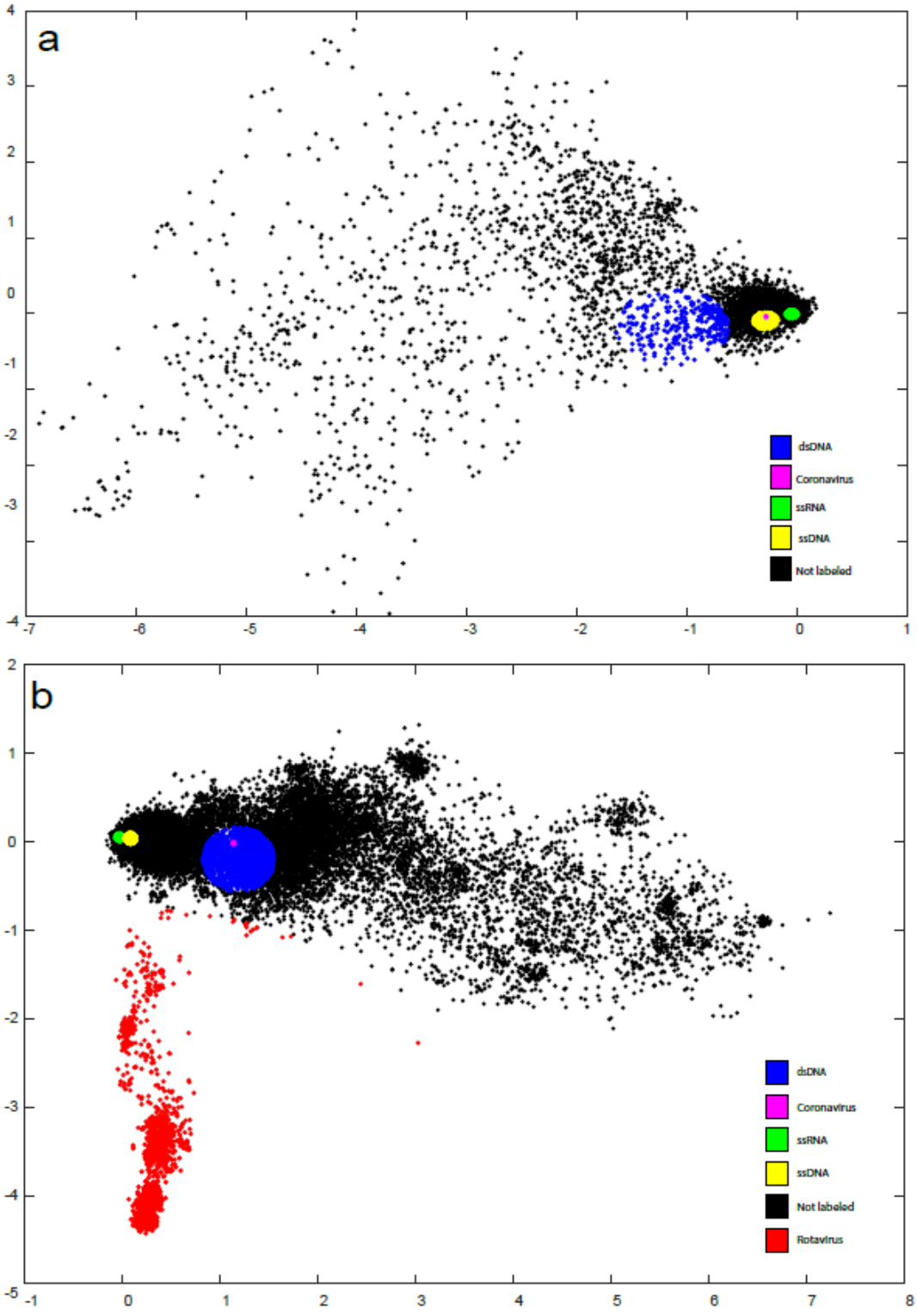


FIGURA 9 – ANÁLISE DOS CONJUNTOS DE DADOS DE VIRAIS

FONTE: Adaptado de FERNANDES *et al*., 2020.

LEGENDA: **a.** PCA do conjunto de dados dos 4.833 proteomas virais: RNA de fita simples (verde), DNA de fita simples (amarelo), DNA de fita dupla (azul), Coronavírus (rosa), não rotulado (preto). **b**. PCA do conjunto de dados dos 31.386 proteomas virais: Espécies com RNA de fita simples (verde), DNA de fita simples (amarelo), DNA de fita dupla (azul), família rotavirdae (vermelha), coronavírus (rosa), não rotulada (preta).

4.4 ANÁLISE DO SARS-COV

De acordo com Gorbalenya e colegas (2020), os critérios adotados para definir o coronavírus responsável pelo surto atual de SARS, como um “novo” coronavírus não seriam os ideais, visto que esta seria a mesma estirpe descrita por Drostren e colegas (2003). De acordo com o ICTV, as nomenclaturas são: SARS-CoV-1, para o vírus isolado em 2003, e SARSCoV-2, para o vírus isolado em 2019, e para um vírus ser considerado uma nova espécie ele não pode estar incluído em nenhum grupo conhecido (GORBALENYA *et al.*; 2020). As análises filogenéticas presentes aqui corroboraram as ideias de Gorbalenya e colegas (2020). Na árvore de 4.833 espécies (FIGURA 9) identificamos que a cepa de pneumonia isolada no mercado de frutos do mar de Wuhan (GCF_009858895.2), responsável pelo atual surto de SARS, está posicionada notavelmente próximo das cepas de coronavírus relacionadas à síndrome respiratória aguda grave (GCF_000864885.1) e coronavírus bat-beta (GCF_000926915.1).

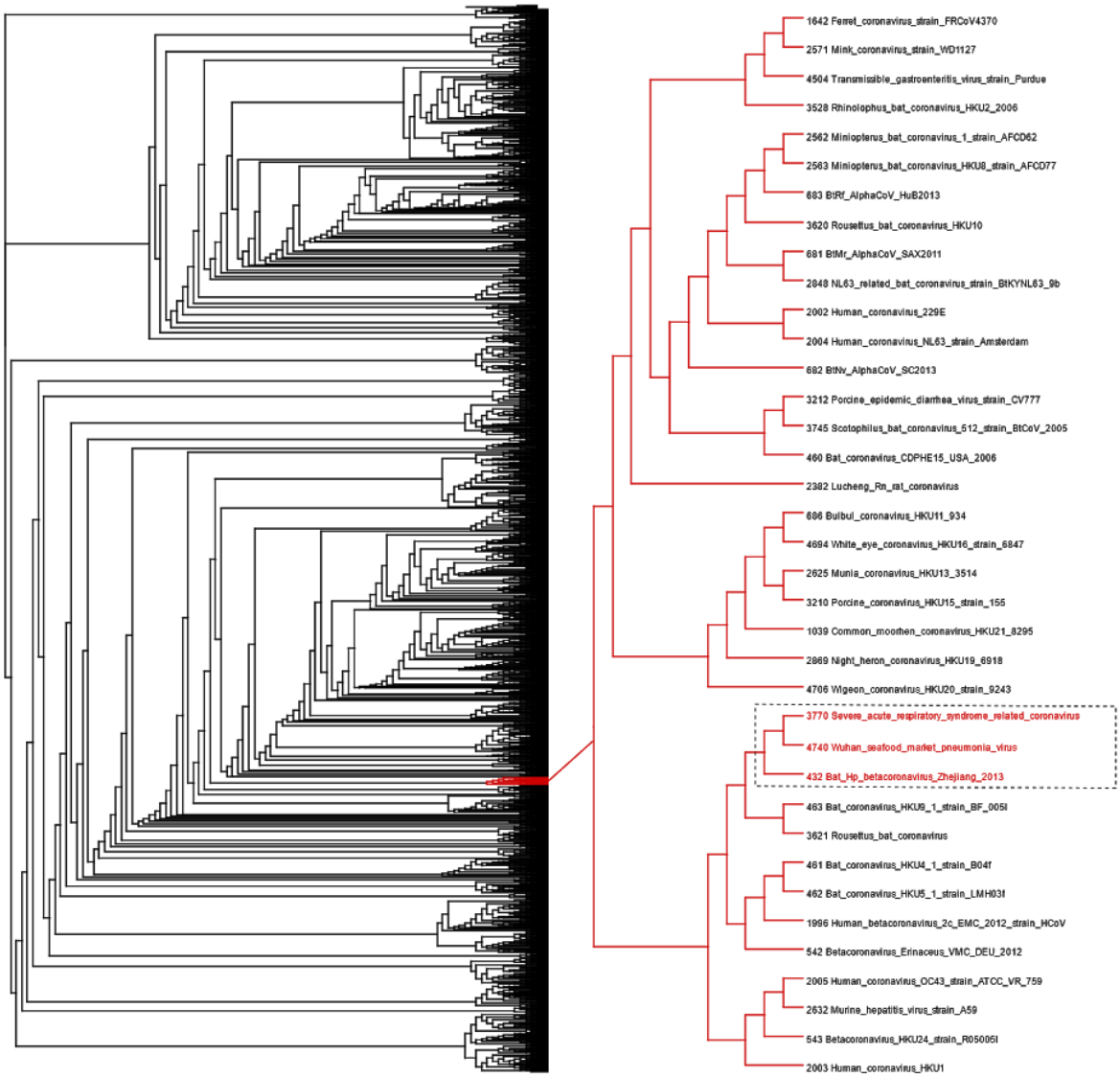


FIGURA 10 - REPRESENTAÇÃO FILOGENÉTICA DE 4,833 PROTEOMAS VIRAIS

FONTE: Adaptado de FERNANDES et al., 2020.
LEGENDA: Árvore filogenética gerada usando dados de 4k, com parâmetros padrão do *SWeeP*: o ramo aumentado contendo as amostras de coronavírus (vermelho). Destaque, as espécies altamente relacionadas à SARS-CoV- 1 (quadrada).

5 CONSIDERAÇÕES FINAIS

Os resultados dos testes de performance mostraram que a implementação do modelo *SWeeP* em linguagem de programação R, o pacote r*SWeeP*, é tao eficiente quanto as implementações originais. As inferências filogenéticas de análise de PCA demonstraram a eficiência do r*SWeeP* para análise e classificação taxonômica de proteomas virais. Desta forma o pacote r*SWeeP* apresenta uma solução para classificação taxonômica de organismos, gratuita e disponível para toda a comunidade científica.

Neste trabalho, novos elementos para uma análise global da filogenia do grupo dos vírus foram descobertos. Além disso, houve a disponibilização das mais completas árvores filogenéticas virais até o momento. As inferências filogenéticas propostas aqui podem apoiar debates relacionados às diferentes estirpes dos vírus SARS-CoV, além de contribuir com a consolidação das ferramentas livre de alinhamento como método de análise de grandes conjuntos de dados biológicos.

Quanto as funcionalidades do pacote r*SWeeP*, futuramente será possível alcançar as possibilidades descritas pela teoria do método *SWeeP*, como a reversibilidade dos vetores *SWeeP* nas sequências originais.

REFERÊNCIAS

1. ABDI, H; WILLIAMS, L. J. Principal component analysis. **Wiley**, v. 2, p. 433-459, 2010.
2. ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool. **Journal of molecular biology**, v. 215, n. 3, p. 403– 10, 1990.
3. ARUNACHALAM, M.; JAYASURYA, K.; TOMANCAK, P.; OHLER, U. An alignment free method to identify candidate orthologous enhancers in multiple Drosophila genomes. **Bioinformatics**, v. 26, n. 17, p. 2109–2115, 2010.
4. BAO, Y. FEDERHEN, S. LEIPE, D. *et al.* National center for biotechnology information viral genomes project. **Journal of virology**, v. 78(14), p. 7291–7298, 2004.
5. BODEN, M.; SCHÖNEICH, M.; HORWEGE, S. Alignment-free sequence comparison with spaced k-mers. **German Conference on Bioinformatics 2013**, v. 34, p. 24–34, 2013.
6. BOOPATHI, S.; POMA, A. B.; KOLANDAIVEL, P. Novel 2019 coronavirus structure, mechanism of action, antiviral drug promises and rule out against its treatment. ***Journal of biomolecular structure & dynamics***, p. 1–10, 2020.
7. CALISHER, C. H. MAHY, B. W. J. Taxonomy: get it right or leave it alone. T**he American Journal of Tropical Medicine and Hygiene**, v. 68, p. 505–506, 2003.
8. CALISHER, C. H.; CHILDS, J. E.; FIELD, H. E. *et al.* Bats: Important Reservoir Hosts of Emerging Viruses Clinical, **Microbiology Reviews**, v. 19 (3), p. 531-545, 2006.
9. CHAPMAN, J. L.; REISS, M. J. **Ecology**: Principles and Applications. Suiça: Ågren, 1992.
10. CUI, J.; LI, F.; SHI, Z. Origin and evolution of pathogenic coronaviruses. **Nature Reviews Microbiology**, v. 17, p. 81–192, 2019.
11. DAVISON, A. J.; LEFKOWITZ, E. J.; SABANADZOVIC, S.; SIMMOND, P. The ICTV Report on Virus Classification and Taxon Nomenclature, 2019. Disponível em: <https://talk.ictvonline.org/ictv-reports/ictv_online_report/introduction> Acesso em: 27 jul. 2020.
12. DE LIMA NICHIO, B. T., DE OLIVEIRA, A., DE PIERRI, C. R., SANTOS, L., LEJAMBRE, A. Q., VIALLE, R. A., DA ROCHA COIMBRA, N. A., GUIZELINI, D., MARCHAUKOSKI, J. N., DE OLIVEIRA PEDROSA, F., & RAITTZ, R. T. RAFTS^{3}G: an efficient and versatile clustering software to analyses in large protein datasets. ***BMC bioinformatics***, *20* (1), 392, 2019.
13. DE PIERRI, C.R. Representações vetoriais de proteomas: um estudo de caso com sequências mitocondriais. Dissertação de Mestrado. Programa de Pós-graduação em Bioinformática, Universidade Federal do Paraná, Curitiba, 2017.
14. DE PIERRI, C.R.; VOYCEIK, R.; RAITTZ, R.T. *et al. SWeeP*: representing large biological sequences datasets in compact vectors. **Scientific Reports**, v. 10(1), p. 91. 2020.
15. DELSUC, F.; BRINKMANN, H.; PHILIPPE, H. Phylogenomics and the reconstruction of the tree of life. **Nature reviews. Genetics**, v. 6, n. 5, p. 361–375, 2005.
16. DROSTEN C.; GÜNTHER S.; PREISER W. *et al*. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. **New England Journal of Medicine**, v. 348(20), p. 1967-1976, 2003.
17. EDGAR, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Research**, v. 32, n. 5, p. 1792–1797, 2004.
18. FERNANDES, D.R.; KULIK, M.G.; MACHADO, D.J.S.; MARCHAUKOSKI, J.N.; PEDROSA, F.O.; DE PIERRI, C.R.; RAITTZ, R.T. rSWeeP: representationA R/Bioconductor package deal with SWeeP sequences. **BioRxiv**, 2020, doi: https://doi.org/10.1101/2020.09.09.290247.
19. GE, J., MENG, J., GUO, N. *et al*. Counting Kmers for Biological Sequences at Large Scale. **Interdiscip Sci Comput Life Sci** 12, 99–108 (2020). https://doi.org/10.1007/s12539-019-00348-5
20. GORBALENYA, A. E., BAKER, S. C., BARIC, R. S. *et al*. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. **Natural Microbiology**, v. 5, p. 536–544, 2020.

21. HOLDER, M.; LEWIS, P. O. Phylogeny estimation: traditional and Bayesian approaches. **Nature Reviews Genetics**, v. 4, n. 4, p. 275–284, 2003.
22. HOLMES, S., HUBER, W. Modern Statistics for Biology. DeNBI: German
23. HORWEGE, S. *et al*. Spaced words and kmacs: Fast alignment-free sequence comparison based on inexact word matches. **Nucleic Acids Research,** v. 42, p. 7–11, 2014.
24. JOHNSON, W. B.; LINDENSTRAUSS, J. Extensions of Lipschitz mappings into a Hilbert space. **Contemp. Math.**, v. 26, p. 189–206, 1984.
25. JUN, S.R.; SIMS, G. E.; W.U., G. A; KIM, S.H. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. Proceedings of the National Academy of Sciences of the United States of America, v. 107, n. 1, p. 133–138, 2010.
26. KARUPIAH, G. **Fields Virology, 4Th Edition**. 2002.
27. Kirchdoerfer R. N.; Cottrell C. A.; Wang N.; Pallesen J.; Yassine H. M.; Turner H. L.; Corbett K. S.; Graham B. S.; McLellan J. S.; Ward A. B. Pre-fusion structure of a human coronavirus spike protein. ***Nature***, v. 531(7592), p. 118–121, 2016.
28. KNIOE, D. M.; HOWLEY, P. M. **Virologia Veterinária**. Santa Maria: UFSM, 2007.
29. LEFKOWITZ, E. J., DEMPSEY, D. M., HENDRICKSON, R. C., *et al*. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). **Nucleic acids research**, 46 (D1), D708–D717 (2018).
30. LEIMEISTER, C. A. BODEN, M. HORWEGE, S. LINDNER, S. MORGENSTERN, B. Fast alignment-free sequence comparison using spaced-word frequencies. **Bioinformatics**. v. 30, p.1991–1999, 2014.
31. LESK, A. M. **Introdução à Bioinformática**, 2 ed. Porto Alegre, RS: Artmed. 348 p, 2008.
32. LI, Y. HE, L. LUCY He, R. YAU, S. S. T. A novel fast vector method for genetic sequence comparison. **Scientific Reports**, 7, p. 1–11, 2017.
33. Maddison, D. R.; Swofford, D. L.; Maddison, W. P. (1997). «NEXUS: an extensible file format for systematic information.». *Systematic biology*. **46** (4). p. 590–621. ISSN 1063-5157. PMID 11975335. doi:10.1093/sysbio/46.4.590
34. MAHMOOD, K.; WEBB, G. I.; SONG, J.; WHISSTOCK, J. C.; KONAGURTHU, A.S. Efficient large-scale protein sequence comparison and gene matching to identify orthologs and co-orthologs. **Nucleic Acids Research**, v. 40, n. 6, 2012.
35. MICHENER, D., SOKAL, R.R. A quantitative approach to a problem of classification. Evolution, v. 11(2), p. 130-162, 1957.
36. MOORE, J. H. Bioinformatics. J**ournal of Cellular Physiology**, v. 213, n. 2, p. 365-369, 2007. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/jcp.21218/full>.
37. MORGAN, M. **Bioconductor Annual Report,** 2019. Buffalo. Disponível em: <https://bioconductor.org/about/annual-reports/AnnRep2019.pdf> Acesso em: 01 jul. 2019.
38. NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Journal of Molecular Biology**, v. 48, n. 3, p. 443–453, 1970.
39. NOÉ, L. MARTIN, D. E. K. A coverage criterion for spaced seeds and its applications to support vector machine string kernels and k-mer distances. **Journal of Computational Biology**, v. 21, p. 947–963, 2014.
40. NORVIG, Peter; RUSSELL, Stuart. Inteligencia artificial. Editora Campus, v. 20, 2004.
41. PEARSON K. On lines and planes of closest fit to systems of points in space. **Philosophical Magazine,** v. (6) 2, p. 559-572, 1901.
42. PELLEGRINI, J.C. **Álgebra Linear**. 223 p., 2016. Disponível em: <http://aleph0.info/cursos/al/notas/al.pdf>
43. RIZZO, J.; ROUCHKA, E. C. Review of Phylogenetic Tree Construction. **University of Louisville Bioinformatics Laboratory Technical Report Series**, 2007.
44. SAITOU N, NEI M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, v.4(4), p.406–425, 1987.
45. SAKAMOTO, T. Ferramentas para análise filogenética e de distribuição taxonômica de genes ortólogos. Tese de doutorado. Universidade Federal de Minas Gerais: Belo horizonte. 2016.

46. SALEMI, M.; VANDAMME, A. **The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny.** Cambridge University Press, Cambridge, UK. 466p, 2003.
47. SANTOS, R.J. **Álgebra Linear e Aplicações**. Universidade Federal de Minas Gerais, 2010.
48. SIMS, G. E.; JUN, S.-R.; WU, G. A.; KIM, S.-H. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. **Proceedings of the National Academy of Sciences of the United States of America**, v. 106, n. 8, p. 2677–2682, 2009.
49. THOMPSON, J. D.; HIGGINS, D. G.; GIBSON, T. J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acids Research**, v. 22, n. 22, p. 4673–4680, 1994.
50. VIALLE, R. A.; PEDROSA, F. D. O.; WEISS, V. A. RAFTS3: Rapid Alignment-Free Tool for Sequence Similarity Search. , p. 1–32, 2016.
51. VINGA, S. Editorial: Alignment-free methods in computational biology. **Briefings in Bioinformatics**, v. 15, p. 341–342, 2014.
52. VINGA, S., ALMEIDA, J. Alignment-free sequence comparison - A review. **Bioinformatics.** v. 19, p. 513–523, 2003.
53. WARD, J. H. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. v.58, p.236–244, 1963.
54. WILKS, S. S. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. **Annals of Mathematical Statistics.** 9: 60–62, 1938.
55. WOLD, S.; ESBENSEN, K.; GELADI, P. Principal component analysis. **Chemometrics and Intelligent Laboratory Systems**, v.2, p. 37-52, 1987.
56. World Health Organization, *Coronavirus Disease 2019 Situation Report 101*. **World Health Organization**, 2020.
57. YANG, Z.; RANNALA, B. Molecular phylogenetics: principles and practice. **Nature Reviews Genetics**, v. 13, n. 5, p. 303–314, 2012
58. ZAHA, A.; FERREIRA, H.; PASSAGLIA L. Biologia Molecular Básica. 5. ed. Porto Alegre; Artmed 2014.
59. ZAK, A. M.; VAN BOHEEMEN, S.; BESTEBROER, T. M.; OSTERHAUS, A.D.M.E.; FOUCHIER, R.A.M. Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. **The New England Journal Of Medicine**,v. 19, p. 1814–1820 , 2012.
60. ZAKI, M. J.; KARYPIS, G.; YANG, J. Data Mining in Bioinformatics (BIOKDD). **Algorithms for molecular biology: AMB**, v. 2, p. 4, 2007.
61. ZHANG, Q. JUN, S. R. LEUZE, M. Ussery, D. NOOKAEW, I. Viral phylogenomics using an alignment-free method: A three-step approach to determine optimal length of k-mer. **Scientific Reports**, v. 7, p. 1–13, 2017.
62. ZIELEZINSKI, A., VINGA S., ALMEIDA J., KARLOWSKI W. M. Alignment-free sequence comparison: benefits, applications, and tools. **Genome Biology**, v. 18, p. 186, 2017.