UNIVERSIDADE DO ESTADO DA BAHIA

DEPARTAMENTO DE CIÊNCIAS EXATAS E DA TERRA

CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

DIEGO DOS SANTOS FONSECA

IDENTIFICAÇÃO DAS REGIÕES GÊNICAS COM MAIOR INFORMAÇÃO FILOGENÉTICA PARA SIMPLIFICAÇÃO DA GENOTIPAGEM IN SILICO

SALVADOR

2021

DIEGO DOS SANTOS FONSECA

IDENTIFICAÇÃO DAS REGIÕES GÊNICAS COM MAIOR INFORMAÇÃO FILOGENÉTICA PARA SIMPLIFICAÇÃO DA GENOTIPAGEM IN SILICO

Monografia apresentada ao curso de Sistemas de Informação do Departamento de Ciências Exatas e da Terra da Universidade do Estado da Bahia - UNEB, como requisito necessário para obtenção do grau de bacharel em Sistemas de Informação.

Área de Concentração: Sistemas de Informação
Linha de Pesquisa: Bioinformática

Orientador: Prof. Ph.D Diego Gervasio Frías Suárez

Orientadora: Prof. Ph.D Maria Inês Valderrama Restovic

SALVADOR

2021

Fonseca, Diego dos Santos
IDENTIFICAÇÃO DAS REGIÕES GÊNICAS COM MAIOR INFORMAÇÃO FILOGENÉTICA PARA SIMPLIFICAÇÃO DA GENOTIPAGEM IN SILICO/ Diego dos Santos Fonseca. – Salvador, 2021
67 p. : il. (algumas color.) ; 30 cm

Orientador: Prof. Ph.D Diego Gervasio Frías Suárez
Orientadora: Prof. Ph.D Maria Inês Valderrama Restovic

1. Genotipagem. 2. Bioinformática. I. Suárez, Diego Gervasio Frías. II. Restovic, Maria Inês Valderrama. III. Universidade do Estado da Bahia. IV. Departamento de Ciências Exatas e da Terra. V. Identificação das Regiões Gênicas com Maior Informação Filogenética para Simplificação da Genotipagem In Silico

DIEGO DOS SANTOS FONSECA

IDENTIFICAÇÃO DAS REGIÕES GÊNICAS COM MAIOR INFORMAÇÃO FILOGENÉTICA PARA SIMPLIFICAÇÃO DA GENOTIPAGEM IN SILICO

Monografia apresentada ao curso de Sistemas de Informação do Departamento de Ciências Exatas e da Terra da Universidade do Estado da Bahia - UNEB, como requisito necessário para obtenção do grau de bacharel em Sistemas de Informação.
Área de Concentração: Sistemas de Informação
Linha de Pesquisa: Bioinformática

Aprovada em:

BANCA EXAMINADORA

Prof. Ph.D Diego Gervasio Frías Suárez (Orientador)
Universidade do Estado da Bahia - UNEB

Prof. Ph.D Maria Inês Valderrama Restovic (Orientadora)
Universidade do Estado da Bahia - UNEB

Prof. Ph.D Ernesto de Souza Massa Neto
Universidade do Estado da Bahia - UNEB

Dedico a minha mãe. Sua grande força foi a mola propulsora que permitiu o meu avanço, mesmo durante os momentos mais difíceis.

Obrigado minha irmã, Débora, pelo apoio em todos os momentos da minha vida.

Agradeço aos meus tios José Barbosa,Antônia Maria, André Rodrigues, Nadir Rodrigues, Sônia Maria e primos pelo apoio e suporte que me deram durante toda a minha vida.

Agradeço a minha companheira, Natália Braga, pelo companheirismo e pela compreensão.

Sou grato meus professores e orietadores Diego Gervasio Frías Suárez e Maria Inês Valderrama Restovic, cuja dedicação e paciência serviram como pilares de sustentação para a conclusão deste trabalho. Grato por tudo.

RESUMO

Após alguns arbovírus surpreenderem o mundo com sua rápida disseminação, o Instituto Evandro Chagas alertou que circulam no território nacional cerca de 210 vírus desta família, sendo que, pelo menos 37 destes são capazes de provocar doenças em humanos. Dentre eles, estão a Dengue, a Chikungunya, e o Zika vírus,sendo que este último, foi encontrado em humanos com doenças febris no oeste da África em 1954 e a partir desta data se espalhou pela Indonésia, Micronésia, Tailândia, Filipina, polinésia francesa, ilha de páscoa e em 2015, pelas Américas. Desde então, começaram os estudos visando conhecer essa espécie e seus genótipos, com o intuito de desenvolver tratamentos eficazes e programas de prevenções para evitar futuros surtos. O procedimento padrão do Sistema de Saúde brasileiro prescreve que, ao identificar-se a presença de um vírus no organismo de um paciente, é necessário reconhecê-lo, determinar a sorotipagem, que consiste na definição do sorotipo ou linhagem do vírus, pois para cada variação é necessário um ou mais tipos de tratamento. Por este motivo, estudam-se as diferenças comportamentais e evolutivas dos genótipos, com o objetivo de identificar e caracteriza-los. Neste trabalho foi realizado o estudo de alguns conjunto de sequência do vírus da Dengue e do Zika Vírus, utilizando um método para a descoberta de regiões com informações filogenéticas e identificação de padrões genéticos nestas mesmas áreas. Os resultados foram favoráveis, pois foram encontradas famílias de padrões que correspondiam aos subtipos dos arbovírus estudados.

Palavras-chave: Dengue.Chikungunya.Zika.Vírus.arbovírus.Filogenética.Genotipagem.Bioinformática

ABSTRACT

After some arboviruses surprised the world with their rapid spread, the Evandro Chagas Institute warned that about 210 viruses of this family circulate in the national territory, and at least 37 of these are capable of causing diseases in humans. Among them are Dengue, Chikungunya, and Zika virus, the latter of which was found in humans with febrile illnesses in West Africa in 1954 and from that date has spread to Indonesia, Micronesia, Thailand, Philippines, Polynesia France, Easter Island and in 2015 for the Americas. Since then, studies have begun to understand this species and its genotypes, with the aim of developing effective treatments and prevention programs to avoid future outbreaks. The standard procedure of the Brazilian Health System prescribes that, when identifying the presence of a virus in a patient's body, it is necessary to recognize it, determine the serotyping, which consists of defining the serotype or strain of the virus, since for each Variation requires one or more types of treatment. For this reason, the behavioral and evolutionary differences of genotypes are studied, with the aim of identifying and characterizing them. In this work, the study of a set of sequences of all subtypes of Dengue and Zika Virus was carried out, using a method for the discovery of regions with phylogenetic information and identification of genetic patterns in these same areas. The results were favorable, as families of patterns were found that corresponded to the subtypes of the arboviruses studied.

Keywords: Dengue.Chikungunya.arboviruses.Zika.virus.phylogenetic.phylogenetic.genotyping

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A descoberta das bases nitrogenadas Adenina, Timina, Guanina e Citosina, associadas aos ácidos nucleicos, DNA (Ácido Desoxirribonucleico) e RNA (Ácido ribonucleico) no ano de 1953 pelos cientistas Watson e Crick, iniciou a história da biologia molecular. Desde então, estuda-se os padrões moleculares baseados em aprofundamento genético e bioquímico (TORTORA et al., 2016). Esses padrões definem as estruturas, as funções do material genético, seus produtos de expressão (proteínas) e analisa a interação entre os diversos sistemas celulares, como o DNA,o RNA e a síntese proteica (ALBERTS et al.,).

As moléculas de DNA contêm as especificações para milhares de proteínas, dado que, cada parte da sequência são transcritas em moléculas de mRNA(RNA mensageiro) separadas, com cada uma codificando uma proteína. Esses segmentos representam o gene responsável por transmitir as características hereditárias à sua prole, incluindo sua forma, seu metabolismo, sua habilidade de se mover, sua capacidade de interagir com outros organismos, entre outros (ZAHA et al., 2014). Esses conjuntos de genes recebidos de herança pelos seus ancestrais são denominados genótipos (ZAHA et al., 2014). Dentre os indivíduos que possuem um vasto número de genótipos, estão os vírus, pois possuem uma taxa de mutação elevada (GERARD et al., 2012).

Uma população de vírus com características similares que ocupam um nicho ecológico específico é definida como espécie viral pelo Comitê Internacional em Taxonomia viral. Essas espécies são agrupadas em gêneros por partilharem características comuns, além de serem designadas por nomes descritivos vulgares, como o Zika Virus, e as subespécies, se existirem, são designadas com um número (Zika – tipo 1). Dentre os gêneros virais existentes, será estudado neste trabalho o Flavivírus (OLIVEIRA et al., 2009).

Os vírus do gênero Flavivírus, são vírus envelopados, de morfologia esférica, com cerca de 50 nm de diâmetro, contendo uma fita simples de ácido ribonucleico (RNA), de polaridade positiva, com aproximadamente 11 kb (ZHANG et al., 2003). Seu genoma é composto por duas regiões não codificantes, denominadas Untranslated Region (UTR), que flanqueiam uma única sequência aberta de leitura (ORF - open reading frame), a qual codifica três proteínas estruturais (capsídeo [C], membrana [M], e envelope [E]) e sete proteínas não estruturais (NS1, NS2a, NS2B, NS3, NS4A, NS4B e NS5), que desempenham funções reguladoras e de expressão

do vírus, como a replicação, a virulência e a patogenicidade (ZAHA et al., 2014).

Entre os arbovírus do gênero citado nos parágrafos anteriores, estão a Dengue, doença que ocorre em áreas tropicais e subtropicais, a Chikungunya, descrito inicialmente na Tanzania em 1950 e o Zika Vírus, que passou a se chamar Zika, porque foi primeiro identificado em macacos rhesus na floresta Zika em Uganda em 1947. O ZIKV foi encontrado em humanos com doenças febris em o oeste da África em 1954. Desde essa data se espalhadou em regiões com a Indonésia, a Micronésia, a Tailândia, a Filipinas e a polinésia francesa, assim como na ilha de páscoa em 2014. A partir de 2015, o vírus começou a se espalhar rapidamente pelas Américas. (ZAHA et al., 2014). Para entender a dinâmica evolutiva deste arbovírus, foi necessário utilizar alguns métodos bioinformáticos para realizar as análises filogenéticas de todas as sequências de alguns arbovírus que continham o gene E, obtidas através do GenBank.

Alinhar as metodologias de pesquisa aos recursos tecnológicos é fator primordial na obtenção de resultados fidedignos no menor tempo e custo possíveis. (KAISER; BENICIO, 2015) Neste ponto, entra a bioinformática, pois está pautada na multidisciplinaridade, envolvendo áreas como biologia molecular, ciência da computação, matemática, estatística e engenharia de softwares. (PROSDOCIMI et al., 2002) Todas com um único propósito: desenvolver novas técnicas capazes de armazenar e analisar dados. Então, a bioinformática é definida como a área de estudo responsável pelo desenvolvimento de ferramentas computacionais para adquirir, armazenar, organizar e analisar dados genéticos(PROSDOCIMI et al., 2002). Dentre as etapas que compõem o processo de análise filogenético estão o sequenciamento, o alinhamento e a construção da árvore filogenética.

A etapa inicial corresponde ao sequenciamento, que é o nome dado ao processo de determinação da ordem sequencial das partes constituintes de um biopolímero não ramificado. Isto é, a ordem de nucleotídeos de uma molécula de DNA ou RNA (KAISER; BENICIO, 2015). O sequenciamento completo de genomas foi possível após avanços tecnológicos que incluem o método de Shotgun e a técnica de Sanger, mecanismos que tornaram possíveis o sequenciamento de fitas longas de DNA. Nos sequenciadores da nova geração a fase de clonagem foi suprimida. Entretanto, nos dois casos, o sequenciamento é feito de maneira aleatória e, em seguida software são utilizados para fazer a sobreposição de sequência, em um processo chamado de montagem (ZAHA et al., 2014). O objetivo desse processo é a obtenção de uma sequência contígua do DNA ou RNA, que será armazenada em um banco de dados genético.

Com a popularidade da internet como força mundial de comunicação, a partir da década de 1980, os bancos de dados biológicos começaram a proliferar e passaram a ser armazenados e administrados por instituições de vários países. Neste trabalho será utilizado o banco de dados primário, GenBank, que é mantido pela *NCBI(National Center for Biotechnology Information)* e contém milhares de sequências de nucleotídeos e aminoácidos armazenadas. Também será utilizado o banco de dados secundário, que é mantido pelo centro de pesquisa em bioinformática da Universidade do Estado da Bahia – UNEB, que é constituído por alguns docentes e discentes do curso de Sistemas de Informação do Campus I, Departamento de Ciências Exatas e da Terra.

Então, ao obter as sequências através dos bancos de dados citados, é necessário alinhá-las. Por isso, utiliza-se o alinhamento, que são técnicas de comparação entre duas ou mais sequências biológicas, que buscam séries de caracteres individuais que se encontram na mesma ordem nas sequências analisadas (VERLI, 2014). Através dos métodos de alinhamento, é possível obter informações a respeito da relação evolutiva entre organismos, indivíduos, genes ou entre sequências diversas. Se duas sequências distintas podem ser alinhadas com certo grau de similaridade, é possível inicialmente assumir que elas compartilharam, em algum momento do tempo passado, um ancestral comum e, por isso, são evolutivamente relacionadas (VERLI, 2014). Como não existe um registro direto das alterações genômicas que as espécies acumularam ao longo dos anos, somente é possível reconstruir o processo de evolução do genoma a partir de comparações detalhadas entre genomas de organismos contemporâneos, e assim construir a árvore filogenética do mesmo. Uma das principais ferramentas utilizadas para este fim, é o *Mega(Molecular Evolutionary Genetics Analysis)* (GERARD et al., 2012).

Conforme vimos anteriormente, o processo de genotipação de um genoma viral é um recurso primordial para a definição do melhor tratamento de um ser infectado por seres acelulares (VERLI, 2014). No entanto, este procedimento é composto de várias etapas que trabalham comparando duas ou mais sequências completas de nucleotídeos ou aminoácidos, com o intuíto de encontrar similaridades entre elas. Então, para sua execução é necessário dispor de alta capacidade computacional.

O objetivo geral deste trabalho foi aplicar um novo método de análise de sequências no estudo de genomas de vírus de importância epidemiológica, visando simplificar o processo de genotipação. Para isso, foram cumpridos o objetivos parciais de desenvolver um método

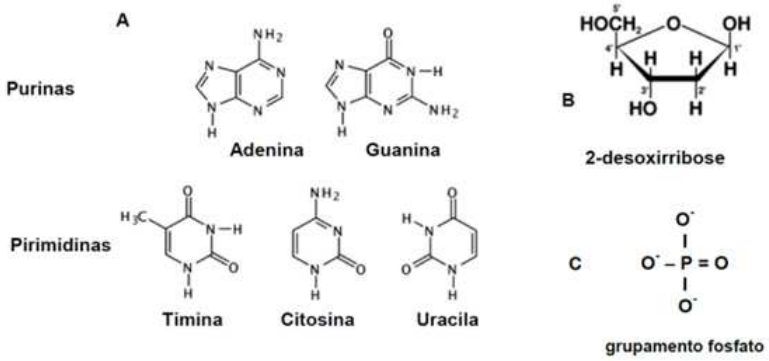
de análise de sequências, para identificação de regiões com maior identificação filogenética e a utilização desta área com identificação filogenética para genotipar um vírus com base nos critérios determinados na metodologia proposta no Capítulo 7 desta monografia.

2 BIOLOGIA MOLECULAR

No início do século XX, foi publicada uma pesquisa que demonstrava pela primeira vez a existência de variação na genética molecular das populações humanas. Este evento despertou o interesse da comunidade cientifica pela área, levando a realização de diversos estudos que culminaram na descoberta da molécula que compunha os genes, o ácido desoxirribonucleico (DNA), e posteriormente foi identificado seu funcionamento e sua estrutura em dupla hélice, por Watson e Crick. A partir destes acontecimentos surgiu a genética molecular (ZAHA et al., 2014).

Do ponto de vista biológico, o DNA é formado de monômeros denominados de nucleotídeos, que são compostos por uma base nitrogenada, um açúcar e um resíduo de ácido fosfórico ligados de forma covalente (Figura 1). Estas bases nitrogenadas podem ser de dois tipos: pirimidinas e purinas. As pirimidinas são a citosina (C), timina (T) e uracila (U), além de apresentarem um anel aromático. Já as purinas são a adenina (A) e guanina (G), e possuem dois anéis aromáticos. O açúcar é uma pentose, a 2-desoxirribose que estabelece uma ligação glicosídica entre o seu carbono C-1'(um linha) e o nitrogênio N-1 das pirimidinas ou o nitrogênio N-9 das purinas, portanto, ligação N-glicosídica. O ácido fosfórico se liga ao carbono C-5' da pentose através de uma ligação éster. O composto formado apenas por uma das bases nitrogenadas e a pentose, ligados de forma covalente, é denominado de nucleosídeo. Nos ácidos nucleicos ocorrem processos que são essências para o funcionamento das células. São eles, a replicação, transcrição e a tradução (GERARD et al., 2012).

Figura 1 – Estrutura da Bases Nitrogenadas

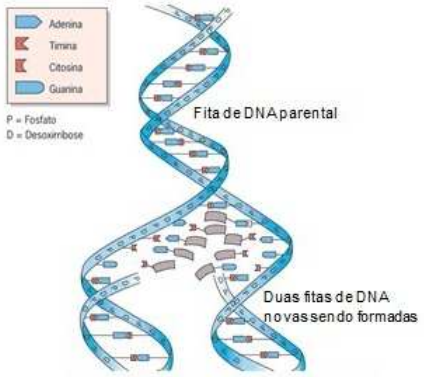


Fonte: (OLIVEIRA et al., 2009).

2.1 REPLICAÇÃO

A replicação é o processo em que cada cadeia de DNA origina duas cadeias filhas semelhantes à ela, através de um processo semiconservativo. Método que conserva nas células resultantes metade das informações genéticas herdadas da célula genitora. Inicialmente, o filamento da cadeia de DNA, tem sua dupla fita separada devido rompimento das pontes de hidrogênio mantidas entre as bases nitrogenadas complementares. Sobre cada uma das fitas molde, vão se emparelhando novos nucleotídeos dispersos no núcleo, construindo dessa forma uma nova cadeia. No final do processo são produzidas duas moléculas idênticas, denominadas de cromátides irmãs, cada constituída por uma fita nucleotídica da molécula original e outra recém-fabricada (Figura 2) (ZAHA et al., 2014).

Figura 2 – Replicação do DNA



5: (ALBERTS et al.,).

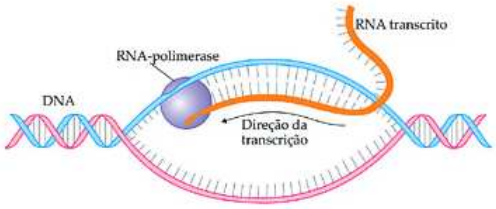
2.2 TRANSCRIÇÃO

Na transcrição, a informação sai do genoma de DNA para a geração de mRNAs (RNAs Mensageiros), que regem toda a maquinaria celular. Então, transcrição é a síntese de uma molécula de ácido ribonucleico (RNA) complementar a um filamento molde de ácido desoxirribonucleico (DNA). Os RNAs produzidos nas células procarióticas e eucarióticas são moléculas de uma única fita composta de nucleotídeos de adenina, de guanina, de citosina e de uracila unida por ligações fosfodiéster que apresentam estruturas secundárias (Figura 3) (ZAHA et al., 2014).

A enzima que atua é a RNA polimerase, que também atua no sentido 5'-3', porém não é necessário a enzima primase para iniciar a polimerização. Como essa enzima não possui atividade revisora, acontece molécula de RNA produzir algumas proteínas defeituosas. Uma

das fitas de DNA aberta serve de molde para a síntese de uma cadeia de RNA mensageiro complementar a fita molde, e que codifica para um gene que será expresso na forma de proteína. Então, os genes possuem regiões que controlam sua própria expressão, os promotores. Os promotores são sequências de nucleotídeos onde se ligam moléculas que inibem ou ativam a transcrição. Serve também, como ponto de ligação de um complexo de proteínas que auxiliam a RNA polimerase a se ligar e agir (ZAHA et al., 2014).

Figura 3 – Transcrição do DNA



Fonte: (ALBERTS et al.,).

2.3 TRADUÇÃO

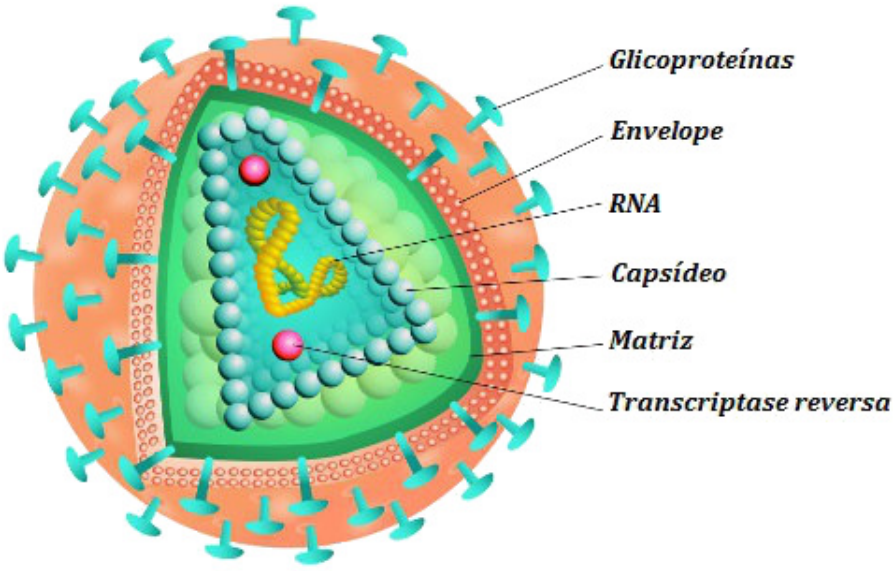
A síntese de proteínas é a etapa final da transferência de informação genética, armazenada no DNA, para as moléculas de proteínas, que são os principais componentes estruturais e funcionais das células vivas. Durante a tradução essa informação, expressa em um RNA, é utilizada para comandar a síntese de uma proteína. O processo de tradução envolve três componentes principais: o RNA mensageiro (RNAm) que contém a informação necessária para direcionar a síntese de proteínas, o RNA de transferência (RNAt) que carregam os aminoácidos que serão incorporados à proteína e os ribossomos que reúnem o RNAm e o RNAt, de modo a permitir que o aminoácido correto seja incorporado à proteína. A tradução começa próximo à extremidade 5', que corresponde ao terminal amino da proteína e prossegue em direção à extremidade 3' do RNA, que corresponde ao terminal carboxila da proteína. A mensagem genética está contida em um código triplo, não sobreposto, sem vírgulas, degenerado e universal. Somente uma combinação das quatro bases existentes no RNA (A, T, C e U) três a três pode gerar o número de combinações ou códons (64) necessários para codificar cada um dos 20 aminoácidos que podem ocorrer nas proteínas. Nenhuma base é compartilhada entre códons consecutivos. O ribossomo move-se ao longo de três bases por vez e como não existe qualquer base interveniente entre os códons, o código é denominado sem vírgulas. O código é degenerado, porque mais de um códon podem codificar o mesmo aminoácido e universal, porque é o mesmo seja em bactérias ou no homem. Três códons (UAA, UAG e UGA) não especificam aminoácido

e são utilizados como sinais para interromper a síntese de uma proteína. O códon AUG, que especifica somente a metionina, tem um duplo papel: ele codifica a metionina em qualquer lugar em que ele se encontre no RNA e também marca o início da síntese proteica (ALBERTS et al.,).

2.4 VÍRUS

Os vírus são pequenas entidades visíveis em sua imensa maioria apenas com microscópio eletrônico, e carregam uma quantidade mínima de material genético constituído por uma ou várias moléculas de DNA ou RNA. Este material é protegido por um envoltório externo ao capsídeo, composto de uma bicamada fosfolipídica e por proteínas imensas nessa bicamada. Os vírus não possuem organização celular, não são capazes de produzir sua própria energia metabólica e precisam necessariamente invadir e controlar uma célula para poder se replicar e dispensar, sendo considerados parasitas intracelulares obrigatórios. Então, os aminoácidos, os nucleotídeos, os ribossomos e a energia metabólica são obtidos a partir de seus hospedeiros. Além disso, diferentemente dos organismos formados por células, os vírus são incapazes de crescer em tamanho e de se dividir autonomamente (ALBERTS et al.,).

Figura 4 – Estrutura Viral



Fonte: (GERARD et al., 2012).

Os vírus não se diferenciam de outros agentes infecciosos somente por serem filtráveis e por requerem células hospedeiras vivas para se multiplicarem, pois essas duas propriedades são compartilhadas por determinadas bactérias pequenas, como as riquétsias. As características

que realmente distinguem os vírus estão relacionadas à sua organização estrutural simples e aos mecanismos de multiplicação. Dessa forma, os vírus são entidades que: Contêm um único tipo de ácido nucleico, DNA ou RNA, contém um invólucro proteico (algumas vezes recoberto por um envelope de lipídios, proteínas e carboidratos) que envolve o ácido nucleico, chamada envelope (Figura 4). Além de utilizar o interior das células vivas para se multiplicar, através da maquinaria de síntese celular e da síntese de estruturas especializadas na transferência do ácido nucleico viral para outras células (TORTORA et al., 2016).

Como os vírus possuem poucas ou mesmo nenhuma enzima própria para seu metabolismo, eles assumem o controle da estrutura da célula hospedeira para se multiplicar. Uma vez no interior destas partículas, estes pequenos seres podem dar origem a alguns ou a milhares de outros seres iguais. Processo no qual, poderá alterar drasticamente a célula, podendo causar sua morte. No entanto, em algumas infecções virais a mesma sobrevive e continua a produzir vírus indefinidamente (OLIVEIRA et al., 2009).

Eles representam uma importante força ecológica, pois parasitam todo tipo de vida na terra, inclusive os micro-organismos, como unicelulares eucariontes, bactérias e arqueias. Possuem grande importância médica e econômica, sendo a causa de dezenas de doenças infecciosas nos seres humanos. Por exemplo, a Síndrome da Imunodeficiência Adquirida(Sigla SIDA, derivada do português, ou AIDS, derivada do inglês), causada pelo vírus HIV é desde a década de 1980 uma das principais causas de morte no planeta e a doença infecciosa que mais atinge pessoas (ZAHA et al., 2014).

Diferentemente da maioria dos organismos que utilizam apenas o DNA (ácido desoxirribonucleico) para armazenar a informação genética, os vírus podem também utilizar o RNA(ácido ribonucleico). A grande variedade de arquiteturas genômicas encontradas nos vírus, é atribuída às diferentes estratégias aplicadas para penetrar nas células. No entanto, essa organização do genoma não determina a letalidade ou a infecciosidade do vírus (GERARD et al., 2012).

2.4.1 Vírus não envelopado

Os vírus mais simples, chamados de não envelopados, são compostos apenas de material genético envolto por um capsídeo proteico. O capsídeo, cápsula protéica, é formado por proteínas produzidas na célula parasitada sob o comando do material genético do vírus. Em

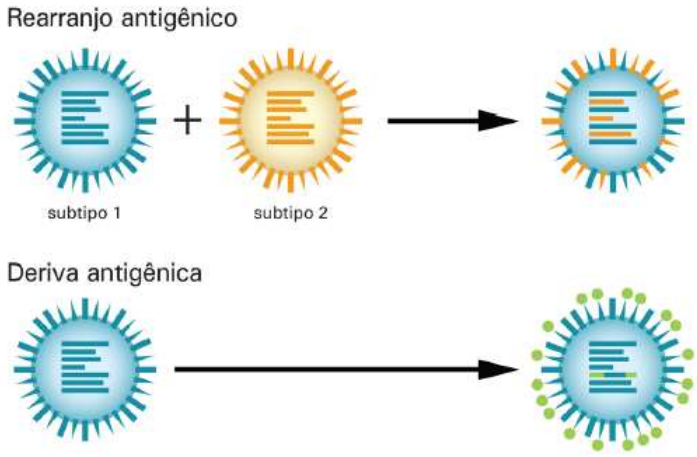
geral, dentro do capsídeo está localizado apenas o material genético, mas podem estar presentes também algumas proteínas que serão essenciais para a invasão da célula hospedeira (OLIVEIRA et al., 2009).

2.4.2 Vírus Envelopado

Alguns vírus podem utilizar parte da membrana da célula hospedeira para recobrir a parte externa do capsídeo. Esses são conhecidos como vírus envelopados. O envelope é, em geral, composto da bicamada fosfolipídica da membrana celular da célula hospedeira, associada as proteínas de origem viral (Figura 4) (ALBERTS et al.,).

2.4.3 Subtipos

Figura 5 – Subtipo Viral



Fonte: (GERARD et al., 2012).

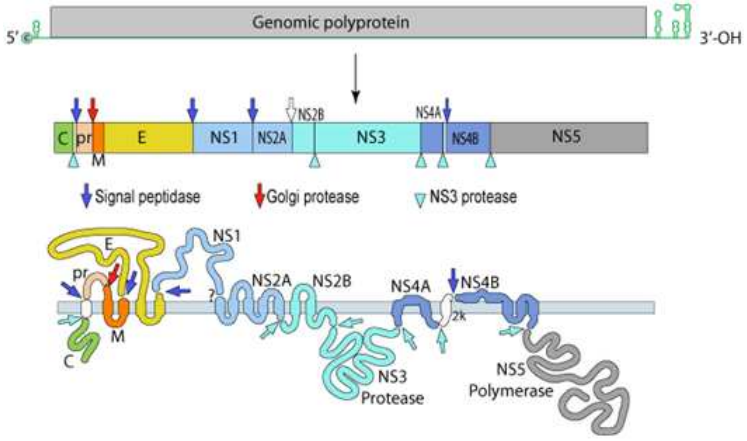
A grande variedade encontrada nessas proteínas de superfície, bem como em todo o material genético viral, provém de mutações e recombinações dos segmentos de RNA que compõem o genoma víral. O acúmulo gradual de mutações que, ao longo do tempo, culmina em versões diferente, porém efetivas, de proteínas de superfície virais é chamado de deriva antigênica. A evolução dessas moléculas tornam os anticorpos produzidos para combater o vírus da gripe ancestral incapazes de reconhecer o novo subtipo viral, tornando-os ineficazes no combate à doença. A deriva antigênica é umas das razões pela qual vacinas para a gripe devem ser desenvolvidas anualmente, a cada temporada de gripe. Cientistas realizam estudos no intuito de prever as mudanças mais prováveis de correr nas proteínas presentes nos vírus que circulam

atualmente no ambiente e, a partir dessa previsão, desenvolver vacinas. Existem portanto, uma probabilidade da vacina produzida ser efetiva, bem como uma probabilidade de não ser, caso a previsão das mudanças não for apurada. Outro processo genômico que ocorre com o vírus é mudança ou redistribuição antigênica (TORTORA et al., 2016).

Neste processo, dois ou mais tipos diferentes do vírus se combinam e formam um vírus totalmente diferente dos ancestrais, criando novos subtipos de hemaglutinina e neuraminidase (Figura 5). Assim, um vírus suíno e outro humano podem se recombinar em uma ave e dar origem a um novo tipo radicalmente diferente de vírus, contendo antígenos completamente desconhecidos para o nosso organismo. Quando a equivalência com outros vírus é menor que 50%, é definido um novo tipo e atribui-se um novo número, na ordem da descoberta. Se a equivalência for maior que 50%, indica-se um novo subtipo e, se próxima de 100%, os vírus são considerados variantes do mesmo tipo. (TORTORA et al., 2016).

2.4.4 Gênero Flavivírus

Figura 6 – Genoma do Flavivírus



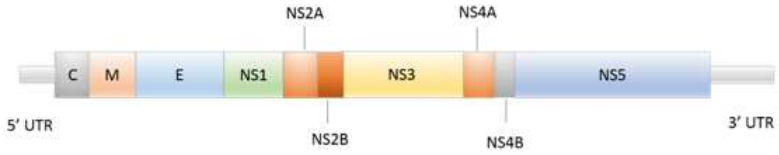
Fonte: (GERARD et al., 2012).

Os vírus do gênero Flavivírus, são vírus envelopados, de morfologia esférica, com cerca de 50 nm de diâmetro, contendo uma fita simples de ácido ribonucleico (RNA), de polaridade positiva, com aproximadamente 11 kb, (ZHANG, CHAPMAN, et al., 2003). Seu genoma é composto por duas regiões não codificantes, denominadas Untranslated Region (UTR), que flanqueiam uma única sequência aberta de leitura (ORF), a qual codifica três proteínas estruturais (capsídeo [C], membrana [M], e envelope [E]) e sete proteínas não estruturais (NS1, NS2a, NS2B, NS3, NS4A, NS4B e NS5), que desempenham funções reguladoras e de expressão

do vírus, como a replicação, virulência e patogenicidade (Figura 6).

2.4.5 Vírus Zika

Figura 7 – Mapa Esquemático do Genoma do Vírus Zika



Fonte: (OLIVEIRA et al., 2009).

O vírus zika (ZIKV) é um arbovírus que pertence ao gênero dos flavivirus dentro da família Flaviviridae igual ao vírus da dengue (Figura 7). O vírus foi chamado de zika, porque foi primeiro identificado em macacos rhesus na floresta Zika em Uganda em 1947. O ZIKV foi encontrado em humanos com doenças febris no oeste da África em 1954. Desde essa data, tem se espalhado pela Indonésia, Micronésia, Tailândia, Filipinas e Polinésia Francesa assim como na Ilha de Páscoa em 2014. O vírus zika se espalhou rapidamente pelas Américas a partir de 2015 (PETERSEN, 2016).

Para entender a dinâmica evolutiva da ZIKV, foram realizadas análises filogenéticas de todas as sequências que continham o gene E e o gene NS5 de cepas ZIKV obtidas do GenBank (YE, LIU, et al., 2016), os resultados desta análise determinaram linhagens do ZIKV como mostra a Tabela 1.

Tabela 1 – Genótipos do Zika Virus.

Vírus	Genótipo
Zika	Africano
	Asiático
	Divergente do Oeste Africano

Fonte: (OLIVEIRA et al., 2009).

2.4.6 Genotipagem

Os métodos de genotipagem estudam o genoma do microrganismo causador da doença, analisando as características do polimorfismo genético dos mesmos (A¡N et al., 2017).

Dividimos os métodos de genotipagem em dois, os métodos in vitro, em que, o processo biológico é realizado em um ambiente fechado e controlado. Já os métodos in silico, são processos realizados através de simulações computacionais.

- Métodos in vitro
  - Reação em cadeia da polimerase (PCR)

    A reação em cadeia de polimerase é um método que amplifica uma única ou poucas cópias de um pedaço de DNA e baseia-se no processo de replicação do DNA que ocorre in vitro(A¡N et al., 2017).
  - Hibridização de sondas de DNA

    A hibridização de sondas é conhecida como a análise em amostras para detectar a presença de ácidos nucleicos (DNA ou RNA), realizando uma combinação antiparalela destes com uma molécula de fita dupla. Suas técnicas são utilizadas para detectar uma molécula alvo a partir de uma sonda complementar a ela(A¡N et al., 2017).
  - RAPD (polimorfismo de DNA amplificado aleatoriamente)

    O RAPD é uma técnica que utiliza marcadores moleculares para amplificação por PCR de sequências curtas de DNA polimórfico usando um trecho curto de sequência (10 a 12 pares de bases)(A¡N et al., 2017).
  - RFLP (polimorfismo de comprimento de fragmento de restrição)

    O RFLP é uma técnica que explora variações em sequências homólogas de DNA, conhecidas como polimorfismos, para distinguir indivíduos, populações ou espécies ou para localizar os genes dentro de uma sequência(A¡N et al., 2017).

- Método in silico
  - Arvore Filogenética

    Árvores filogenéticas são estruturas que expressam a similaridade, ancestralidade e relacionamentos entre as espécies ou grupo de espécies. Conhecidas como árvores evolucionárias ou simplesmente filogenias, as árvores filogenéticas possuem folhas que representam as espécies (táxons) e nós internos que correspondem aos seus ancestrais hipotético(VIANA, 2007).São necessárias três etapas para construir uma árvore filogenética a partir de sequências de nucleotídeos ou aminoácidos(RITTER et al., 2019).
    * Alinhamento das sequências;

* Escolha de um método de substituição de nucleotídeos ou aminoácidos adequado;
* E escolha do método de reconstrução filogenética;

3 BIOINFORMÁTICA

A bioinformática é uma ciência Multidisciplinar, pois envolve a engenharia de softwares, a matemática, a física, a química, a estatística e a ciência da computação com o intuito de compreender as funções biológicas, mais especificamente os genes (VERLI, 2014). Essa ciência é responsável pela aquisição, análise e armazenamento de informações biológicas sob a forma de ácido nucleicos e de proteínas, com o auxílio de métodos computacionais e de algoritmos matemáticos. Assim, reconhece padrões que provavelmente seriam impossíveis de serem analisados sem tal ajuda. Como consequência direta da grande quantidade de informações que são geradas, diversos banco de dados genéticos foram criados.

3.1 BANCO DE DADOS GENÉTICOS

Na década de 80 a internet se popularizou e como força mundial de comunicação, e neste período foram criadas redes em biologia molecular nos EUA e Europa, com o objetivo de atender a demanda por informações sobre área. Então os bancos de dados começaram a se proliferar e tiveram que ser administrados por instituições de vários países. Desde então, surgiram vários tipos de banco de dados que foram agrupados de acordo com as informações que armazenam. Esses conjuntos de sequências de nucleotídeos, de aminoácidos ou de estruturas de proteínas podem ser classificados em bancos de dados primários e secundários (PROSDOCIMI et al., 2002).

3.1.1 Banco de Dados Primário

Os bancos de dados primários são mantidos por três instituições: *National Center for Biotechnology Information (NCBI)*, o *European Bioinformatics Institute* e o *DNA Data Bank of Japan*. Sendo que, os três bancos compartilham todas as informações. Essas três instituições, junto com o *Universal Protein Resource(UniProt)*, que é um repositório central de sequências de proteínas curadas, administram os maiores repositórios de sequencias de nucleotídeos e de proteínas do mundo. (PROSDOCIMI et al., 2002).

Nesta pesquisa foi utilizado o Genbank, banco de dados mantido pelo *National Center for Biotechnology Information (NCBI)* desde 1992, pois contém uma coleção de sequencias de nucleotídeos e suas traduções de proteínas. O GenBank tornou-se uma base de dados importante

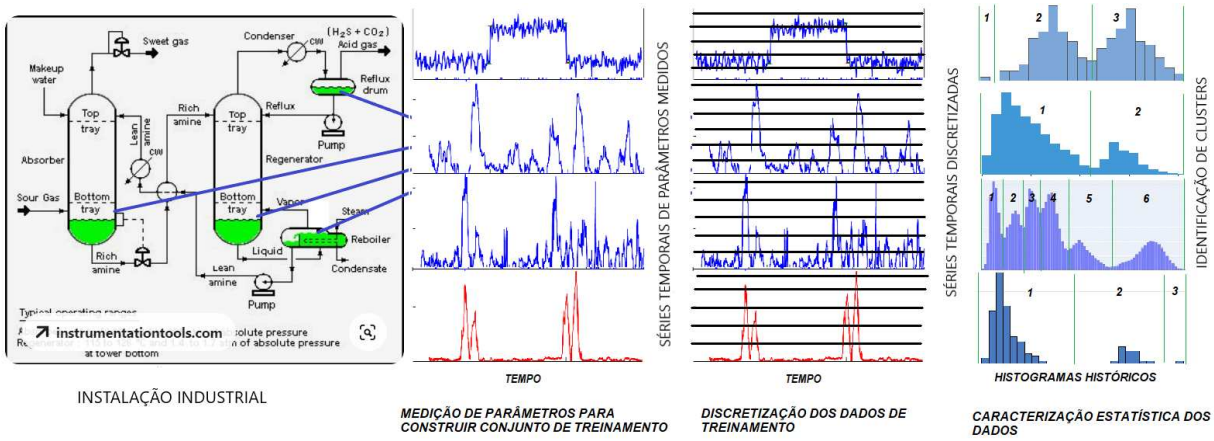
para pesquisa em campos biológicos e cresceu nos últimos anos, com uma taxa exponencial, duplicando aproximadamente a cada 18 meses. (PROSDOCIMI et al., 2002)

3.1.2 Banco de Dados Secundário

Os bancos de dados secundários são aqueles que derivam dos primários, ou seja, foram formados usando as sequências de nucleotídeos ou aminoácidos depositados nos bancos de primários. Nesta pesquisa usamos o ABVdb (Base de dados de vírus Arthropode Borne), banco de dados que contem dados filogenéticos, epidemiológicos e clínicos de sequências de arbovírus,com a Dengue e a Zika Vírus. Mantido pelo centro de pesquisa em bioinformática da Universidade do Estado da Bahia – UNEB, que é constituído por alguns docentes e discentes do curso de Sistemas de Informação do Campus I, Departamento de Ciências Exatas e da Terra.

4 *MÉTODO DE REPRESENTAÇÃO DE ESTADOS PARAMÉTRICOS (PARAMETRIC STATE REPRESENTATION METHOD - PSRM)*

Figura 8 – PSRM (Parametric State Representation Method)



Fonte: Imagem do autor.2021

O método de representação de estados paramétricos (Figura 8), chamado PSRM de acordo com a sigla do termo em inglês, foi desenvolvido para caracterizar dinamicamente o estado dos objetos (sistemas, subsistemas e componentes do sistema de geração e hidroviário da concessionária AES Tietê) supervisionados por sistemas inteligentes de controle de instalações industriais de funcionamento contínuo. O estado é composto por características primárias e secundárias. As características primárias utilizadas para a caracterização dos estados são classificadas de duas formas: (1) Valores das m variáveis medidas pelo sistema de instrumentação e controle e-ou manualmente, associadas ao objeto medido, e (2) Notificação de estados dos $N_c \geqslant 0$ pontos que integram o objeto monitorado. Em total usamos $p = m + N_c$ características primárias. O método estabelece uma ordem fixa das p características primárias, respeitando a ordem das mesmas nos arquivos de configuração, e atribuindo uma escala para a normalização de cada uma delas(Figura 8). Isto pode ser feito manualmente para algumas variáveis, mas também pode ser feito por aprendizado automático, processando bancos de dados históricos utilizados para treinamento dos agentes. A escala utilizada para as características associadas ao estados dos objetos integrantes é fixa para todos os objetos e depende do método utilizado para gerar essa notificação de estado. No projeto, utilizamos os resultados da propagação da perda de potencial de disponibilidade(PAL), que é uma medida que varia no intervalo unitário. Por isso, a escala para esse tipo de característica primária é dispensada.

As caraterísticas secundárias, derivadas das primárias, foram classificadas em dois tipos: (1)Dinâmicas e (2) Correlacionadas. As m características secundárias dinâmicas foram calculadas com a diferença relativa de cada variável medida em dois distantes de amostragens consecutivas. As diferenças são relativas porque se divide a diferença pelo valor médio das variáveis primárias nos dois instantes. Por outro lado, as características secundárias correlacionadas foram calculadas como as diferenças entre todos os pares de variáveis primárias medidas em cada instante de amostragem, o que resulta em $m*(m-1)/2$ características secundárias correlacionadas. Desta forma temos $s = m + m*(m-1)/2$ variáveis secundárias e o estado paramétrico dos objetos monitorados composto por $f = N_c + m*(2+(m-1)/2)$ características ou variáveis de estado.

5 TRABALHOS CORRELATOS

Na procura de uma possível solução para o problema que este trabalho busca solucionar, uma pesquisa foi realizada com a finalidade de encontrar pesquisas científicas que tenham relação com o processo de genotipação de um vírus. Após análise dos 272 artigos encontrados, foi possível perceber que o processo de genotipação de um genoma viral é essencial na tomada de decisão do tratamento mais eficaz para combater esses seres acelulares. Porém os resultados dessa revisão mostram que os métodos desenvolvidos para simplificar esta técnica não se aplicam ao Zika Vírus e nenhum identifica a região genômica correspondente ao genótipo do Vírus.

O artigo “A Simple and Reliable Strategy for BK Virus” - (MOREL et al., 2017), buscou uma estratégia eficaz para caracterizar todas as estirpes do vírus BK, utilizando sequencias completas, arvores filogenéticas, alinhamentos e polimorfismo isolados. Tiveram como resultado, a identificação 12 diferentes subtipos. Este artigo aproxima-se do presente trabalho por propor a construção de uma ferramenta para a identificação de subtipos. Contudo, diferencia-se pelo método adotado e pelo vírus estudado.

Já o artigo “The effect of phylogenetic signal reduction on genotyping of hepatitis E viruses of the species Orthohepevirus A” - (PURDY; SUE, 2017), examina a redução do sinal filogenético da espécie Orthohepevirus A, à medida que o cumprimento das sequencias diminuem. O mesmo aproxima-se do presente trabalho, por estudar regiões com sinais filogenéticos, entretanto, distancia-se no mapeamento destas regiões para futuras genotipagens.

6 METODOLOGIA

Neste capítulo apresenta-se a metodologia adotada no desenvolvimento do método de identificação de regiões com maior informação filogenética, e posteriormente, o método que utilizou essas áreas encontradas, para simplificar o processo de genotipação.

Para melhor entendimento dos métodos que foram desenvolvidos, remetemos o leitor à figura 9, na qual esboçamos todas as fases do projeto.

Figura 9 – Etapas do projeto

Fonte: Imagem do autor.2021

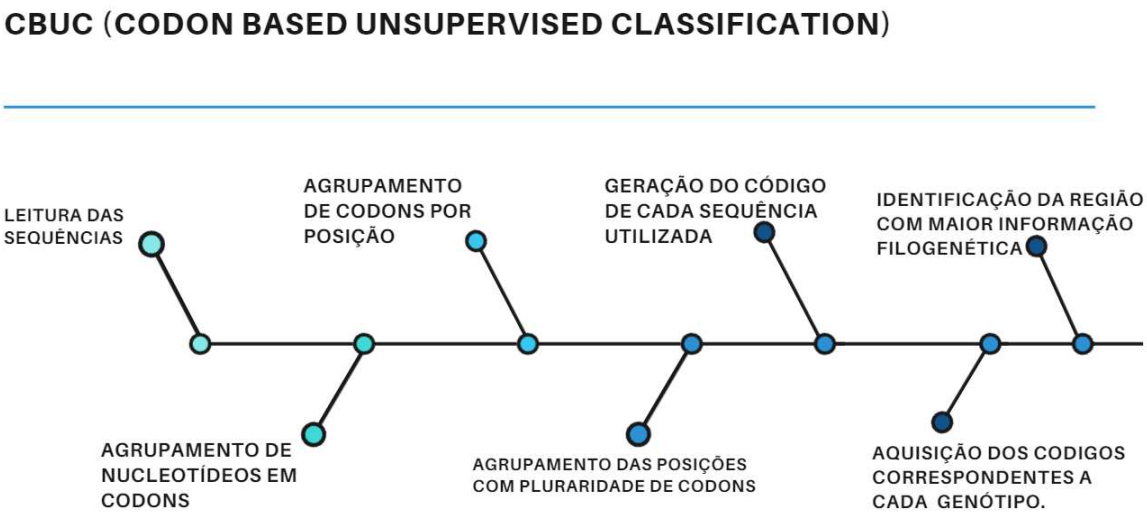
6.1 TRATAMENTO DE DADOS

Na etapa de tratamento de dados, foram coletadas 50 sequências genéticas de cada subtipo do vírus da Zika, através do banco secundário ABVdb. Após obter este conjunto de dados, foi necessário utilizar o software BioEdit Sequence para alinhar todas as cadeias genéticas e cortá-las com intuito de deixa-las com tamanhos semelhantes.

6.2 DESENVOLVIMENTO DO CBUC

Como mostra na Figura 10, inicialmente, foram identificados todos os nucleotídeos da sequência, separando-os em códons (sequência de três bases nitrogenadas que codificam um aminoácido) e atribuindo uma etiqueta de 3 letras e um número do 1 ao 64 ao aminoácido codificado por esse códon, segundo o código genético mostrado na Tabela 2.

Figura 10 – Fases do método CBUC



Fonte: Imagem do autor.2021

Tabela 2 – Tabelas de Códons

n	Codon	aa	n	Codon	aa	n	Codon	aa	n	Codon	aa
1	AAA	Lys	17	GAA	Glu	33	CAA	Gln	49	UAA	STOP
2	AAG		18	GAG		34	CAG		50	UAG	
3	AAC	Asn	19	GAC	Asp	35	CAC	His	51	UAC	Tyr
4	AAU		20	GAU		36	CAU		52	UAU	
5	AGA	Arg(2)	21	GGA	Gly	37	CGA	Arg(4)	53	UGA	STOP
6	AGG		22	GGG		38	CGG		54	UGG	Trp
7	AGC	Ser(2)	23	GGC		39	CGC		55	UGC	Cys
8	AGU		24	GGU		40	CGU		56	UGU	
9	ACA	Thr	25	GCA	Ala	41	CCA	Pro	57	UCA	Ser(4)
10	ACG		26	GCG		42	CCG		58	UCG	
11	ACC		27	GCC		43	CCC		59	UCC	
12	ACU		28	GCU		44	CCU		60	UCU	
13	AUA	Ile	29	GUA	Val	45	CUA	Leu(4)	61	UUA	Leu(2)
14	AUG	Met	30	GUG		46	CUG		62	UUG	
15	AUC	Ile	31	GUC		47	CUC		63	UUC	Phe
16	AUU		32	GUU		48	CUU		64	UUU	

Fonte: (ALBERTS et al.,).

Como observado na Tabela 2, existem 64 códons, sendo deles 61 codificantes (que codificam algum dos 20 aminoácidos existentes) e 3 são códons de terminação da tradução da proteína codificada pelo gene cuja sequência estamos analisando.

Como resultado desta fase, uma sequência com n nucleotídeos (n múltiplo de 3) é compactada a uma sequência de $m = n/3$ códons, representada como uma lista de m inteiros no intervalo de 1 a 64, $S = s(1), s(2), ..., s(m)$, onde $s(j)$ é o número do códon no seu sítio $j = 1, 2, ..., m$.

Varrendo então o dataset, uma sequência de cada vez, se calculada a quantidade (ou a frequência) $Q(p,i)$ (ou $F(p,i)$) de cada um dos códons, $i = 1, 2, ..., 64$, em cada uma das m posições $p = 1, 2, ..., m$, na sequência compactada. Se temos ns sequências no dataset de treinamento, a máxima quantidade $Q(p,i)$ possível de um códon i em uma posição p é igual a ns, pelo que dividindo a quantidade por ns obtém-se a frequência $F(p,i) = Q(p,i)/ns$, que varia entre 0 e 1. O caso $Q(p,i) = ns(F(p,i) = 1)$ acontece nas posições p onde um mesmo códon i está presente nas ns sequências do dataset. Desde o ponto de vista biológico, essas posições são chamadas de "sítios conservados", ou seja, que não hão sofrido mutação durante o processo de diferenciação genética das espécies virais e não possuem valor classificatório, pois o códon é o mesmo para todas as sequências.

A contagem da quantidade $Q(p,i)$ ou da frequência $F(p,i)$ de códons distintos $i = 1, 2, ..., 64$ em uma posição p no intervalo $[1, m]$ é chamada de "histograma". Imagine um gráfico com 64 colunas, uma para cada códon codificante na tabela X, no qual se coloca em cada coluna uma barra de altura igual à quantidade de vezes que o códon correspondente foi encontrado na posição em estudo. Adotando a quantidade Q como variável de trabalho em vez da frequência F, no caso dos sítios conservados, teremos uma única barra de altura ns, ou seja, igual à quantidade de sequências no dataset de treinamento. Já nos sítios não conservados teremos mais de uma coluna com distintas alturas, cumprindo-se que a soma das alturas será igual a ns, ou seja, $Q(p,1) + Q(p,2) + ... + Q(p,64) = ns$.

Pode acontecer que as colunas aparecem separadas ou juntas, ou uma mistura desses dois casos. Por isso introduzimos o conceito de clusters (agrupamento em inglês). Duas ou mais barras juntas, independentemente das suas alturas, formam um cluster composto, da mesma forma que uma barra isolada constitui um cluster simples. Contudo, pode-se assumir que cada barra é um cluster, independentemente se encontra junto ou não com a outra barra. A escolha

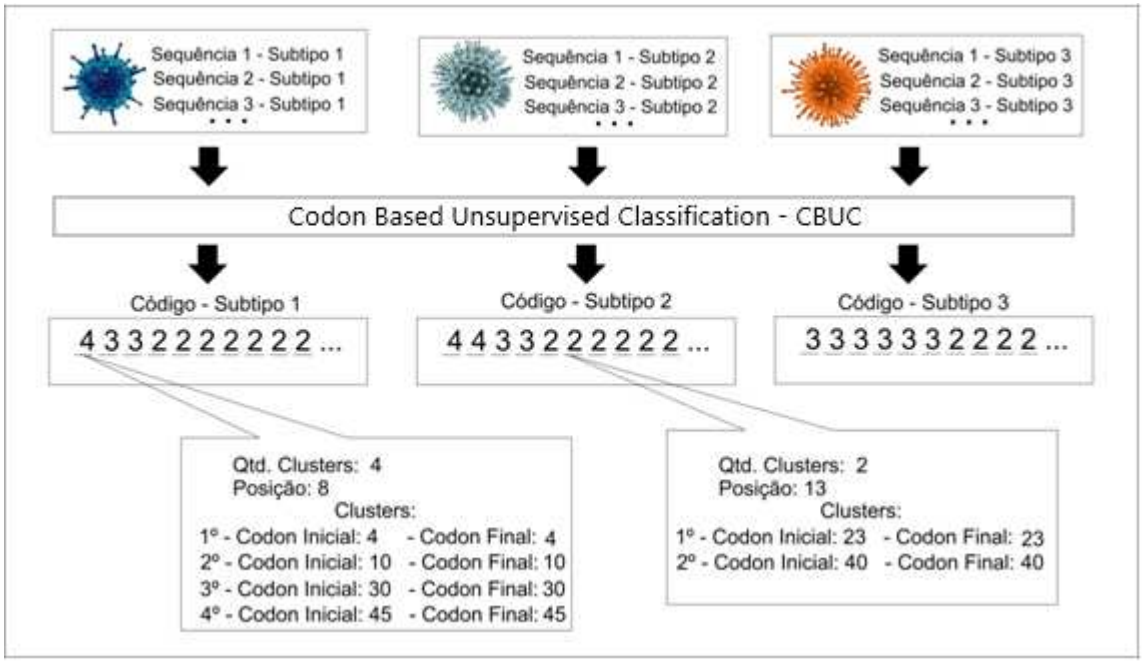
entre estas duas formas de definir clusters é controlada com um parâmetro que chamamos de distância mínima entre clusters, *dmin*. Se definimos $dmin = 1$, então duas barras juntas podem constituir dois clusters diferentes, já que a distância entre elas é de uma coluna que é a mínima estabelecida por dmin. Agora, definindo $dmin > 1$ duas barras juntas não podem constituir dois clusters diferentes porque se encontram a uma distância de uma coluna, menor que dmin e então elas formam parte de um mesmo cluster. Por exemplo, para o caso $dmin = 2$, para que uma barra seja considerada pertencente a outro cluster é necessário que exista pelo menos 1 coluna "vazia" entre ela e a barra (cluster) mais próxima à direita e à esquerda dela. Porém, definindo $dmin = 3$, uma barra que está próxima a outra barra ou cluster, havendo uma coluna vazia entre elas, será considerada parte desse cluster próximo, mesmo havendo um espaço entre eles. A mudança nos resultados de classificação em função do valor de dmin escolhido, foi um dos objetos de estudo neste trabalho.

Uma vez feita a identificação dos clusters em uma posição $p = 1,2,...,m$ usando o valor de *dmin* definido, se registra a quantidade de clusters existentes, $nc(p)$, assim como o número da primeira e última colunas (códons) em cada cluster, $first(c,p)$, $last(c,p)$ para cada cluster $c = 1,2,...,nc(p)$, em cada posição p da sequência de códons. Note que no caso de clusters formados por uma única barra tem-se $first(c,p) = last(c,p)$, ou seja o códon inicial e final são iguais.

O conjunto das variáveis formadas por $nc(p)$, $first(c,p)$ e $last(c,p)$ para toda posição $p = 1,2,...,m$ nas sequências do dataset de treinamento, é chamado de "Mapa de Clusters - MC". O Mapa de Clusters é de fato o conjunto de "atributos classificatórios" ou "features" extraído dos dados de treinamento, além de servirem de base tanto para o processo de identificação das posições e regiões génicas "menos conservadas" (com "maior informação filogenética"), quanto para o processo de classificação não supervisionada das sequências, descritos a seguir.

Na figura 11 mostramos o processo de construção do MC descrito.

Figura 11 – Processo de Construção do Mapa de Clursters



Fonte: Imagem do autor.2021

6.2.1 Identificação das Regiões com Maior Informação Filogenética

Introduzindo um índice de informação filogenética por sítio (códon, aminoácido) dado por:

$$IIF(p) = nc(p) - 1$$

É possível identificar trechos de sequências (intervalos de posições) formados apenas por sítios conservados aonde $IIF(p) = 0$, assim como poderão ser visualizados intervalos com altos valores de IIF que correspondem as regiões com maior informação filogenética, ou seja, as mais mutadas na amostra de treinamento que geralmente correspondem as mais mutáveis.

Outra aplicação importante do IIF é que permite construir um ranking dos sítios do gene de acordo com seu conteúdo de informação filogenética, sendo possível então construir uma sequência sintética curta contendo apenas os sítios com IIF acima de certo valor definido, $IIFmin$. Por exemplo, denotando por $IIFmax$ o máximo valor de IIF na amostra, o $IIFmin$ pode ser definido como uma fração $(percentual/100)f < 1$ do $IIFmax$, ou seja, $IIFmin = f * IIFmax$. Desta forma o tamanho da sequência sintética irá diminuindo na medida que f aumenta. Contudo, pode se aplicar outro critério de escolha do conjunto de sítios mais informativos que fixe o tamanho L da sequência sintética desejada. Neste caso, os sítios devem ser ordenados de maior a

menor segundo seu IIF, ou seja, constrói-se um ranking, do qual se selecionam os primeiros L sítios para compor a sequência sintética representativa desse gene na população viral em estudo.

O uso de sequências sintéticas com alto valor filogenético reduz o custo computacional e pode aumentar a confiabilidade (reduzir a dependência dos modelos e parâmetros utilizados) das inferências filogenéticas atuais. Porém, a avaliação do impacto do uso de sequencias sintéticas está fora do escopo deste trabalho.

Por outro lado, o ranking de sítios por informação filogenética pode servir de guia para o desenho de vacinas e para o estudo dos mecanismos de escape a terapias antivirais.

6.2.2 Classificação não Supervisionada

O método PSRM gera um código w de c dígitos para cada sequência. Duas sequências com igual código pertencem à mesma classe numa primeira instância. Ou seja, o método PSRM gera um número C de classes, cada uma possuindo um único código. Métodos para agrupamento de códigos muito similares em famílias podem ser aplicados para reduzir o número de classes.

O processo de geração do código w começa com o ranking descrito na seção anterior. Considere que a lista de c posições ranqueadas $pr(1), pr(2), ..., pr(c)$ correspondem as posições p no intervalo $[1, m]$ na sequência original resultado do ordenamento da lista $nc(p)$ do Mapa de Clusters, de maior a menor. Como o menor número de clusters nc, é 1 nos sítios não informativos, no início da lista ordenada estarão os sítios com maior número de clusters e no final da lista os sítios não informativos com $nc = 1$. Então, cortando os k sítios do final com $nc = 1$, restam $c = m - -k$, sítios com $nc > 1$. Observe que c define o "maior comprimento" de um código w, o qual permite classificar toda a amostra, mas não necessariamente o "comprimento mínimo". A definição do comprimento mínimo não foi objeto deste estudo.

O código para uma sequência de m códons $S = s(1), s(2), ..., s(m)$, onde $s(j)$ é o número do códon no seu sítio $j = 1, 2, ..., m$, é gerado identificando o códon $j = s(pr(i))$ na i-ésima posição do ranking $pr(i)$, para $i = 1, 2, ..., c$, e depois identificando em qual dos $nc(pr(i))$ clusters o códon j está contido. O ordinal k no intervalo $[1, nc(pr(i))]$ especificando o cluster de pertença é adicionado ao código w na posição i, ou seja, $w(i) = k$, sendo k o inteiro que satisfaz a condição $first(k, pr(i)) jlast(k, pr(i))$.

Vale a pena ressaltar que a quantidade de códigos w diferentes (classes) que pode ser gerada não depende apenas do tamanho c do código, mas vem dada por

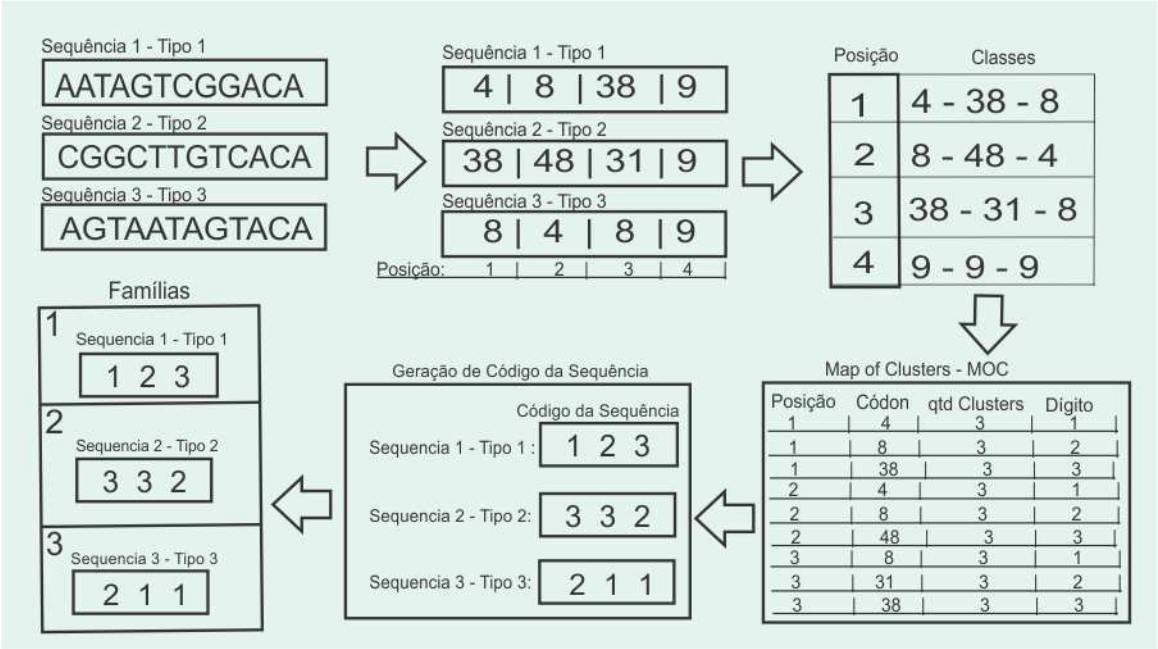
$$C_{max} = \prod_{i=1}^{c} nc(i)$$

Isto por sua vez permite introduzir um índice de variabilidade para as amostras de diferentes espécies virais, definido como:

$$v = c\frac{C}{C_{max}}$$

Na figura 12 mostramos todo o processo do CBUC.

Figura 12 – Comportamento do CBUC



Fonte: Imagem do autor.2021

6.2.3 Identificação do genótipo

Antes do método ser capaz de identificar o genótipo de uma sequência de entrada, é preciso associar os códigos atribuídos às sequências de cada genótipo durante a fase de classificação não supervisionada descrita acima, ao genótipo correspondente. Isto pode ser feito de forma automática utilizando adicionando a cada sequência do dataset de treinamento uma etiqueta com o genótipo. No caso de G genótipos a serem identificados, denotados por $g1, g2, ..., gG$, após serem identificadas C classes de forma não supervisionada, às quais foram

atribuídos os códigos $w1, w2, ..., wC$, se realiza uma associação do subconjunto de códigos que foram atribuídos a cada um dos genótipos. Seja Wg o subconjunto de códigos atribuídos ao genótipo $g[1,G]$ e Cg a quantidade de códigos nesse subconjunto.Uma rápida checagem da unicidade da atribuição é feita verificando se $\sum_{g=1}^{G} C_g = C$. No caso que $\sum_{g=1}^{G} C_g > C$ isto significa que há pelo menos um código que foi atribuído a mais de um genótipo. Mais especificamente indica que há $\sum_{g=1}^{G} C_g - C$ códigos redundantes. Os códigos repetidos em dois genótipos diferentes $g1$ e $g2$ são os que estão na interseção dos respectivos subconjuntos $Wg1$ e $Wg2$. As sequências que geram esses códigos não poderão ser genotipadas de forma não ambígua usando este método, podendo-se ainda dar como resultado os vários subtipos possíveis, em vez de nenhuma saída nesses casos. No caso ideal, todas as sequências de um mesmo genótipo recebem um único código, o que equivale a identificar tantas classes como genótipos,$C = G$. Foi testada a influência de dmin no número e diversidade dos códigos gerados para cada genótipo, assim como o efeito de reduzir o comprimento do código eliminando posições no final do mesmo nos resultados da classificação. Para avaliar a qualidade da genotipagem foram usadas 50 sequências de cada genótipo que não foram incluídas no conjunto usado para treinar o método. Foi calculada a matriz de confusão K de tamanho GxG contendo no elemento da linha i e da coluna j, ou seja, em $k(i,j)$ a quantidade de sequências do genótipo gi, que foi classificada como o genótipo gj. A figura 13 mostra a matriz de confusão K.

Figura 13 – Matriz de Confusão

		GENÓTIPO ATRIBUÍDO			
		g1	g2	...	gG
GENÓTIPO REAL	g1	k(1,1)	k(1,2)	...	k(1,G)
	g2	k(2,1)	k(2,2)	...	k(2,G)
	⋮	⋮	⋮	...	⋮
	gG	k(G,1)	k(G,2)	...	k(G,G)

Fonte: (OLIVEIRA et al., 2009).

As predições corretas se encontram na diagonal da matriz de confusão, $k(i,i)$, para $i = 1,2,...,G$ enquanto os elementos fora da diagonal, $k(i, j \neq i)$, para i , $j = 1,2,...,G$, correspondem às classificações erradas. A métrica de qualidade usada foi definida como:

$$q = max(0, \frac{\sum_{i=1}^{G} k(i,i) - \sum_{i=1}^{G} \sum_{j=1, j \neq i}^{G} k(i,j)}{N_v}) \in [0,1]$$

que avalia a capacidade classificatória em excesso com respeito a um classificador

aleatório, que em média vai acertar 50% das classificações e vai errar também 50%. Observe que se o número de predições corretas (soma na diagonal) for igual ao número de predições incorretas (soma fora da diagonal) então q = 0, o que indica que o classificador é tão bom quanto um classificador aleatório, ou seja, que não possui nenhuma capacidade classificatória adicional. No caso que erra mais do que acerta o valor de q é assumido 0, evitando-se valores negativos que carecem de sentido prático numa métrica de desempenho.

O denominador Nv é a quantidade de sequências no dataset de validação que também é igual à soma de todos os elementos da matriz K. Note que no caso de não haver classificações incorretas todos os elementos ficam na diagonal e sua soma é igual a Nv, pelo que a métrica toma seu máximo valor$q = 1$. A métrica introduzida é válida para amostras balanceadas, ou seja, que contem aproximadamente a mesma quantidade de sequências de cada genótipo.

6.3 GERAÇÃO DO DATASET

Para o armazenamento de todas as informações que foram usadas neste projeto, criamos um banco de dados relacional, fundamentado no paradigma de orientação de conjuntos, e que tem como linguagem padrão o SQL (Structure Query Language). Por atender aos requisitos citados acima, ser gratuito e de código aberto, foi escolhido o banco de dados Mysql.

6.4 RECURSO UTILIZADOS

A seguir serão descritos em detalhes os recursos utilizados no desenvolvimento deste projeto.

6.4.1 Softwares

- NetBeans IDE

  O NetBeans é um ambiente de desenvolvimento integrado gratuito e de código aberto para desenvolvumento de software nas linguagens Java, JavaScript, HTML5, entre outra.
- MySQL

  O MySQL é um sistema de gerenciamento de banco de dados (SGBD) escrito em C e C++, que utiliza a linguagem SQL (Linguagem de Consulta Estruturada, do inglês Structured Query Language) como interface.

- Biblioteca Javasci (Scilab)

  O Javasci é uma biblioteca oriunda do scilab, que permite a execução de alguns comandos e scripts deste mesmo software.
- MEGA - *Molecular Evolutionary Genetics Analysis*

  O MEGA é um software livre disponível para ajudar cientistas e estudantes na elaboração de dendrogramas ou árvores filogenéticas usando sequências de nucleótidos ou proteínas.

6.4.2 Linguagens

- Java

  Java é uma linguagem de programação orientada a objetos desenvolvida na década de 90, que diferentemente das das linguagens de programação modernas, é compilada para um bytecode para ser posteriormente interpretado por uma máquina virtual (Java Virtual Machine, mais conhecida pela sua abreviação JVM).
- SQL - Linguagem de Consulta Estruturada

  A SQL é a linguagem de pesquisa declarativa padrão para banco de dados relacional.
- CSS

  A CSS é É uma linguagem de estilo utilizada para definir a apresentação de documentos escritos em uma linguagem de marcação.
- HTML

  A HTML consiste em uma linguagem de marcação utilizada para produção de páginas na web.

6.4.3 Servidor

- Apache

  O servidor Apache é responsável por disponibilizar páginas e todos os recursos de um website, tais como, envio de e-mails, mensagens e diversas outras funções. O mesmo processa arquivos escritos em diferentes linguagens de programação, como Java, Python e outras.

6.4.4 Padrão de projeto

- MVC

  O MVC é um padrão de arquitetura de software responsável pela separação do projetos em três partes(Model,View e Controller), com o intuito de reduzir suas dependências.

7 *MÉTODO DE AGRUPAMENTO NÃO SUPERVISIONADO BASEADO EM CÓDONS (CODON BASED UNSUPERVISED CLASSIFICATION - CBUC)*

Neste trabalho, o CBUC que foi desenvolvido para processar múltiplas sequências de códons, similares (pertencentes ao mesmo gene-espécie) contidas num conjunto de treinamento.

Na nossa abordagem voltada para a análise de sequências de códons, em vez de nucleotídeos, é feito um "paralelo conceitual"entre um sinal monitorado (de entrada) e a posição de um códon numa sequência de DNA em estudo. A variável temporal do modelo PSRM para séries temporais é portanto equivalente à lista de sequências alinhadas no conjunto de treinamento.

O mesmo mecanismo mediante o qual o PSRM atribui um código único ao estado do sistema em cada instante de tempo partir dos valores de todas as séries temporais monitoradas, é aplicado no método CBUC atribuindo-se um código a cada sequência no conjunto de treinamento.

A partir da similaridade dos códigos é feito o agrupamento (clustering) das sequências do conjunto de treinamento.Então, uma vez definidos os agrupamentos existentes no conjunto de treinamento pelo método CBUC, cada agrupamento é associado por especialista humano ao tipo/subtipo correspondente, como é habitual na classificação não supervisionada. A partir deste momento, ou seja na fase pós-treinamento, qualquer sequência, previamente alinhada, pode ser tipada/subtipada pelo método CBUC. A descrição detalhada de cada passo do método é feita nas seções seguintes.

Além do método CBUC, são descritas duas formas distintas de utilizar a informação contida no modelo de dados criado: (1) Para construir a sequência sintética mais informativa e (2) para extrair a subsequência menor e mais informativa. Estes dois tipos de sequências são relevantes para estudos de filogenia molecular, e nunca antes tinham sido estudadas.

7.1 CONSTRUÇÃO E TREINAMENTO DO MODELO PARA CLASSIFICAÇÃO NÃO SUPERVISIONADA BASEADA EM CÓDONS - CBUC

7.1.1 Extração de Característica de Conjuntos de Dados de Sequência FASTA

A primeira etapa visa transformar as sequências de nucleotídeos de entrada, no formato FASTA, em uma série de atributos inteiros, variando de 1 a 61, indicando o códon

em cada posição na sequência de códons.Para fazer isso, primeiro considere um conjunto de sequência de codificação FASTA de DNA/RNA alinhado em quadro $\mathscr{S} = \{S_1, S_2, \ldots, S_{N_s}\}$, contendo N_s sequências de N_n nucleotídeos (removendo nucleotídeos não identificados "N", lacunas e códons de parada).

Considere um esquema de numeração de nucleotídeo bi-unívoco arbitrário $\mathscr{N}$, que atribui um ordinal $n \in [1,4]$ para cada nucleotídeo $r = \{A, G, C, T \ ou \ U\}$:

$$n = \mathscr{N}(r) \tag{7.1}$$

Portanto, as sequências FASTA no conjunto de dados de entrada $\mathscr{S}$ podem ser compiladas para uma sequência de $N_c = N_n/3$ códons numéricos: (n_1, n_2, n_3), onde n_k é o ordinal do nucleotídeo na k -ésima posição do códon ($k = 1,2,3$), calculado usando o operador de compilador de caractere para inteiro $\mathscr{N}$ na Eq. 7.1.

Considere agora um esquema de numeração de códon bi-unívoco arbitrário $\mathscr{M}$ que atribui um ordinal $c = 1, 2, \ldots, 61$ a cada códon de sentido, isto é

$$c = \mathscr{M}(n_1, n_2, n_3) \tag{7.2}$$

Desta forma, as sequências de códons numéricos $s_i, \ i = 1, 2, \ldots, N_s$ obtidas por uma compilação anterior com o operador $\mathscr{N}$ no nível de nucleotídeo, vamos denotá-lo como

$$s_i = \mathscr{N}(Y_i)$$

são novamente compilados para uma sequência de ordinais de N_c códon:

$$\sigma_i = \mathscr{M}(y_i) = \mathscr{M}(\mathscr{N}(Y_i)) = \{c_1, c_2, \ldots, c_{N_c}\}_i, \ i = 1, 2, \ldots, N_s \tag{7.3}$$

Cada uma das entradas inteiras discretas N_c de σ_i é uma característica primária de nosso modelo. Este conjunto de atributos é posteriormente reduzido (otimizado) conforme descrito abaixo.

7.1.2 Fase de Exploração

Nesta fase, criamos uma lista

$$L_c = \left\{c_1, c_2, \ldots\right\}_c \tag{7.4}$$

com os diferentes códons $c_i,\ i = 1,2,\ldots$ encontrados em cada posição de códon $c = 1,2,\ldots,N_c$, onde c_i é calculado usando a Eq.7.2. Então, adicionamos códons à lista L_c de acordo com a ordem de aparição no conjunto de sequências. Em outras palavras, sempre que um novo códon (não listado em L_c) é encontrado em qualquer posição de códon c, ele é adicionado à lista L_c.No entanto, os códons na lista podem ser classificados de qualquer outra maneira, por exemplo, de acordo com a frequência do códon naquela posição no conjunto de dados de treinamento.

Independentemente da ordem do códon nas listas L_c, nós construímos a lista de listas

$$\mathscr{L} = \{L_1, L_2, \ldots, L_{N_c}\} \tag{7.5}$$

que chamamos de Mapa de códons (MOC) e é a informação do modelo "primária"coletada do conjunto de sequências de treinamento.Em seguida, calculamos o número de códons distintos em cada códon c do local, denotando-o por $\lambda_c = tamanho(L_c),\ c = 1,2,\ldots,N_c$. Observe que se c é uma posição de códon onde foram encontrados diferentes códons no conjunto de treinamento, então $\lambda_c > 1$, e em caso contrário, $\lambda_c = 1$.

A coleção

$$\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_{N_c}\} \tag{7.6}$$

são as informações "secundárias"de modelo coletadas do conjunto de dados de treinamento.

Sabendo o número de códons diferentes, λ_c, em cada posição de códon $c-1,2,\ldots,N_c$, podemos reescrever a lista de códons em cada posição de códon, L_c definido na Eq. 7.4, de uma forma mais detalhada como:

$$L_c = \left\{c_1, c_2, \ldots, c_{\lambda_c}\right\}_c \tag{7.7}$$

7.1.2.1 Conteúdo de Informação Classificatória - CIC

Apresentamos o conteúdo de informação classificatória (CIC), denotado como $\mathscr{I}$, de um códon-site $c = 1,2,\ldots,N_c$ as

$$\mathscr{I}_c = \lambda_c - 1$$

Em outras palavras, o conteúdo de informação do códon de um posição de códon é dado pelo número de códons diferentes encontrados naquela posição.

Antes de continuar é importante estabelecer o paralelo existente entre o conteúdo de informação classificatória (CIC - $\mathscr{I}$) e a intensidade de sinal filogenético utilizado no jargão da Bioinformática (PURDY; SUE, 2017): CIC e sinal filogenético são conceitos equivalentes.

7.1.3 Seleção de atributos

Nossa seleção de atributos é baseada em *CIC*. Formalmente, o ordinal $c_j \in [1.61]$ do códon encontrado no sítio do códon $j \in [1, N_c]$ da sequência de entrada S_i, definido em 7.3, é o valor da j -ésima característica primária do modelo para a sequência de entrada S_i.No entanto, como os locais de códon conservados têm $\mathscr{I} = 0$, classificamos os locais de códon de valores superiores para $\mathscr{I}$ inferiores e descartamos os locais de códon com $\mathscr{I} = 0$.

Para representar esse processo, apresentamos a lista decrescente baseada em *CIC* de sítios de códons com $\mathscr{I} > 0$, como

$$\mathscr{F} = \left\{ p_j \in [1, N_c] \mid \mathscr{I}_{p_j} > 0 \text{ e } \mathscr{I}_{p_j} \geq \mathscr{I}_{p_{j+1}} \text{ para } j \in [1, F] \right\} \tag{7.8}$$

onde F é o número de características selecionadas, isto é, o número de locais de códons que carregam informações classificatórias ($\mathscr{I} > 0$) no conjunto de dados de sequência FASTA de treinamento. Observe que $\mathscr{F}$ acima é apenas uma lista de posições de códons informativos no conjunto de sequências alinhadas, que denotamos aqui como p_j, para $j = 1, 2, \ldots, F$, ordenados de acordo com seus valores $\mathscr{I}$. O vetor $\mathscr{F}$ é a informação do modelo "terciário"coletada do conjunto de dados de treinamento.Do ponto de vista prático, uma vez que o modelo de classificação é construído e o conjunto de atributos $\mathscr{F}$ é definido, apenas os códons nas posições $p_j,\ j \in [1, F]$ na consulta as sequências de entrada serão coletadas e usadas para sua classificação.

Observe também que $\mathscr{F}$ de fato carrega três peças de informação valiosas: (1) o número de locais de códons informativos no conjunto de dados de treinamento (referência), (2) a posição de cada códon informativo, e (3)) a relevância de cada posição do ponto de vista classificatório.

Usamos $\mathscr{F}$ para construir versões compactas e ordenadas do mapa de códons $\mathscr{L}$ (Eq. 7.5) e λ (Eq. 7.6), denotamos como

$$\mathscr{L}^* = \{ L_{p_j}, j \in [1, F] \} \tag{7.9}$$

$$\lambda^* = \{\lambda_{p_j}, j \in [1,F]\} \tag{7.10}$$

onde os códons p_j são aqueles selecionados em $\mathscr{F}$. Observe que λ_{p_j} é o comprimento da lista de códons L_{p_j}, contendo os códons encontrados no local do códon p_j do conjunto de sequências de treinamento alinhado.

7.1.4 Atributos derivados para rotulagem de sequências

Neste ponto, somos capazes de atribuir um rótulo (ou ID ou assinatura) a cada sequência FASTA $S_i,\ i \in [1,N_s]$ do conjunto de dados de treinamento e, claro, a qualquer sequência fora de dados após treinamento, composto de valores de atributos de F.

Seguindo nossa abordagem, os valores dos atributos são naturalmente inteiros, o que do ponto de vista computacional faz com que esse tipo de rótulo seja um vetor discreto (vetor) de F entradas inteiras, a ser usado para calcular uma distância significativa entre dois desses rótulos. Do ponto de vista da ciência dos dados, a escolha de feições inteiras implica na opção por traços categóricos, ao invés dos numéricos, o que traz o conhecido problema em aberto de definir distâncias entre vetores de feições categóricas (REFs). Nossa abordagem para resolver esse problema em nosso contexto de aplicativo é descrita nas próximas seções.

Desconsiderando o problema de definição de distância com características categóricas, pensamos em duas maneiras de definir os inteiros F que formam nosso vetor de características. Uma maneira é usar o número do códon, $c_{p_j},\ j \in [1,f]$, variando de 1 a 61, em cada posição j em $\mathscr{F}$ (veja Eq. 7.2). No entanto, como a numeração do códon é arbitrária e não necessariamente carrega qualquer informação biológica significativa, preferimos usar a posição (ordinal k) do códon c_{p_j} encontrado na posição do códon p_j, para $j = 1,2,\ldots,F$ da sequência, na lista de códons $L_{c_{p_j}}$ definidos na Eq. 7.7, ou seja:

$$k_{c_{p_j}} = \{k \mid c_{p_j} = c_k,\ c_k\ cL_{c_{p_j}}\} \tag{7.11}$$

onde $k \in [1,\lambda_{c_{p_j}}]$ é a posição (ordinal) do códon c_{p_j} na lista $L_{c_{p_j}}$ preenchida de acordo com a Eq. 7.7.

Nossa escolha de atributo inteiro, que é computacionalmente mais caro do que usar o códon ordinal encontrado em cada local do códon, obedece a um critério de visualização mais

curto das marcas das sequências. Embora os ordinais do códon variem entre 1 e 61 e sejam palavras de 2 dígitos, espera-se que o ordinal dos diferentes códons encontrados em cada posição do códon na sequência seja inferior a 10 na maioria dos casos, mesmo em vírus, e então será requerem apenas um caractere a ser escrito.

Portanto, o rótulo (assinatura, etiqueta) $\mathscr{T}_i$, da i -ésima sequência FASTA é composta por um vetor com F ordenado de um dígito $k_{c_{p_j}},\ j = 1,2,\ldots,F$, ou seja

$$\mathscr{T}_i = \{k_{c_{p_1}}, k_{c_{p_2}}, \ldots, k_{c_{p_F}}\}_i, \tag{7.12}$$

para cada sequência FASTA $S_i,\ i = 1,2,\ldots,N_s$ no conjunto de dados de treinamento. Na Eq. 7.12 cada entrada $k_{c_{p_j}},\ j \in [1,F]$ varia no intervalo discreto $[1,\lambda_{c_{p_j}}]$, lembrando que $\lambda_{c_{p_j}} > 1,\ \forall j \in [1,F]$ é o número de códons diferentes nos locais de códons selecionados para o conjunto de dados de treinamento fornecido.

7.1.5 Descoberta da Diversidade

Considerando a capacidade de atribuir uma etiqueta F dígitos $\mathscr{T}$ a cada uma das sequências N_s no conjunto de dados de treinamento S, varremos o conjunto de dados registrando e numerando as diferentes etiquetas. Deixe T denotar a coleção de etiquetas de F dígitos

$$T = \{\mathscr{T}_1, \mathscr{T}_2, \ldots, \mathscr{T}_D\} \tag{7.13}$$

onde D é o número de etiquetas diferentes encontradas no conjunto de dados de treinamento, que reflete a diversidade de padrões de uso de códons, ou seja, D é o número de diferentes padrões de uso de códons encontrados no conjunto de dados de treinamento.

Durante o processo de descoberta de padrões, cada sequência no conjunto de dados é rotulada com o ordinal de sua marca na lista de marcas T. Ao pesquisar um grande treinamento datasets, a taxa de descoberta de novas etiquetas/padrões começa muito alta, com uma nova etiqueta por sequência, mas diminui monotonicamente sendo muito baixa no final do conjunto de dados, pois a maioria das etiquetas nas últimas sequências já foram encontradas antes e estão contidas em T. Por esta razão, na maioria dos casos, exceto em conjuntos de dados não redundantes, o número de padrões encontrados, D, é menor que o número de sequências, N_s, no conjunto de dados de treinamento, pois as sequências redundantes recebem o mesma marca.

7.1.6 Agrupamento das sequências de DNA/RNA no conjunto de treinamento segundo à semelhança da respectivas sequências de códons

A última etapa do nosso método consiste em agrupar as etiquetas encontradas $\mathscr{T}_d,\ d = 1,2,\ldots,D$, não as sequências. Denotando por M_C o número de agrupamentos presentes nos dados e por $\mathscr{C}_i$ o i -ésimo agrupamento, o problema consiste em distribuir as etiquetas D em T (Eq. 7.13) entre M_C agrupamentos, comprovado que $M_C < D$.

O agrupamento é feito agrupando sequências com padrões "semelhantes", que são etiquetas em nosso caso. A similaridade é medida como inversamente proporcional a alguma medida de distância entre padrões no espaço de atributos dimensional F.

Nossos vetores de características são etiquetas D compostas por características discretas (categóricas) F. Vamos reescrever a Eq. 7.12 no formulário

$$\mathscr{T}_d = \{f_{1,d}, f_{2,d}, \ldots, f_{F,d}\},\ d = 1,2,\ldots,D \tag{7.14}$$

onde $f_{k,d} \in C_k,\ k = 1,2,\ldots,F$ é a categoria da k -ésima característica da d -ésima etiqueta (padrão), sendo C_k o subconjunto de categorias para o atributo k.

De acordo com a estrutura e a natureza de nossos vetores de características (Eq. 7.14), usamos a distância de Hamming (WAGGENER, 1995) para medir a distância entre dois pares de etiquetas de mesmo comprimento ($\mathscr{T}_i$ e $\mathscr{T}_j$), denotamos como $\delta_{i,j} = dist\left(\mathscr{T}_i, \mathscr{T}_j\right)$, para $i, j \in [1,D]$.

A distância de Hamming entre duas etiquetas $\mathscr{T}_i$ e $\mathscr{T}_j$, denotada por $H_{i,j}$, é o número de características com diferentes categorias na etiqueta par, de modo que $H_{i,j} \leq D$ em todos os casos. Aqui usamos distâncias normalizadas $\delta_{i,j} = H_{i,j}/D\ \in [0,1]$, para construir uma diagonal estritamente superior $D\ timesD$ matriz de distância:

$$\Delta = \{\delta_{i,j},\ j > i,\ \text{e}\ \delta_{i,j} = 0\ \forall j \leq i,\ \forall i \in [1,D]\}$$

para o conjunto de dados de treinamento.

Mesmo tendo a matriz de distância Δ, é possível aplicar algoritmos de agrupamento hierárquico usuais, como o Agrupamento de Vizinhos (Neighbour Joining - NJ) (SAITOU; NEI, 1987) para construir um dendrograma (uma árvore filogenética em Bioinformática), aqui abordamos o classificação direta das sequências virais de acordo com seu tipo/subtipo, usando métodos de agrupamento não supervisionados.

No entanto, mesmo que para os métodos de agrupamento não supervisionados mais usados, como K-means (MACQUEEN, 1967), o número de agrupamentos, M_C, seja um hiperparâmetro de configuração, optamos por não definir o número esperado de agrupamentos para avaliar a capacidade natural de identificação de agrupamento do método CBUC, mantendo-o assim um método sem parâmetros.

7.1.6.1 Algoritmo de identificação de agrupamentos não supervisionado sem parâmetros

Implementamos uma versão nossa do algoritmo de agrupamento aglomerativo hierárquico ingênuo de link único (F.; P., 2012) com uma complexidade de tempo de $\mathscr{O}(D^2)$. No entanto, também é possível escolher um método diferente.

O método começa com uma lista ordenada

$$\mathscr{P} = \left\{(i,j)_1, (i,j)_2, \ldots, (i,j)_{(D^2-D)/2}\right\}$$

de $(D^2-D)/2$ pares de etiquetas $(i,j),\ j > i, i, j \in [1,D]$, ordenados pelas distâncias $\delta_{i,j}$, de menor a maior, contidas na matriz de distâncias Δ. A saída do método é o "mapa de agrupamentos", que é uma lista $\mathscr{A} = \{A_1, A_2, \ldots, A_G\}$, contendo as listas de etiquetas pertencentes a cada um dos G agrupamentos encontrados nos dados. As listas de etiquetas $A_g,\ g = 1, 2, \ldots, G$ são preenchidas enquanto são percorridos os pares da lista $\mathscr{P}$, de acordo com os seguintes passos:

1. Inicialização:
   - Inicialize o primeiro agrupamento A_g para $g = 1$ com as 2 etiquetas no primeiro par de etiquetas (i_1, j_1) da lista $\mathscr{P}$, ou seja:

     $$A_1 = \{i_1, j_1\}$$

     e atualize o valor do número de agrupamentos $G = 1$ e do cursor p que percorre os pares na lista $\mathscr{P}$, ou seja, $p = 1$.
2. Incremente o cursor, $p = p + 1$, para processar o próximo par de etiquetas em $\mathscr{P}$, que é $(i,j)_p$, e veja se alguma das etiquetas neste par, i_p ou j_p, já está listada em algum dos G agrupamentos encontrados até o momento. Uma das três situações a seguir pode acontecer:
   - Criar um novo agrupamento: Se nenhuma das etiquetas i_p, j_p está contida em algum dos G agrupamentos encontrados, então crie um novo agrupamento incrementando $G = G + 1$, e atribua ambas etiquetas ao novo agrupamento A_G, ou seja:

     $$A_G = \{i_p, j_p\}$$

- Expandir um agrupamento: Se uma das etiquetas (i_p ou j_p), já está contida em um dos G agrupamentos encontrados até o momento, seja o agrupamento $g \in [1,G]$, mas a outra etiqueta não está, então adicione a etiqueta que não está à lista de etiquetas do g-ésimo agrupamento, ou seja, à A_g.
- Mesclar dois agrupamentos: Se uma etiqueta é encontrada em um determinado agrupamento $g_1 \in [1,G]$ enquanto a outra etiqueta é encontrada em um agrupamento diferente $g_2 \in [1,G]$, mescle os agrupamentos A_{g_1} e A_{g_2} reescrevendo o agrupamento A_{g_1} com as duas listas, ou seja $A_{g_1} = A_{g_1} \cup A_{g_2}$, exclua o agrupamento A_{g_2} do mapa de agrupamentos $\mathscr{A}$ e atualize o número de grupos $G = G - 1$.

3. Retorne ao ponto 2 até $p = (D^2 - D)/2$.

Ao final do processo de agrupamento, cada etiqueta foi atribuída a um dos G agrupamentos encontrados, o que permite atribuir um agrupamento a cada uma das sequências do conjunto de dados de treinamento, de acordo com a etiqueta da sequência. Desta forma, é possível avaliar o desempenho da classificação não supervisionada fazendo: (1) A associação de cada agrupamento com um tipo/subtipo de vírus correspondente, se houver correspondência entre o número de agrupamentos, G, e o número de tipos/subtipos no conjunto de dados de treinamento, e (2) O cálculo de taxas de predição corretas e erradas e métricas de desempenho multi-classe apropriadas.

Nas próximas seções descrevemos o uso do modelo para:

1. Classificar sequências novas (não contidas no conjunto de treinamento, de acordo com seu tipo/subtipo viral
2. Construir sequências sintéticas contendo apenas os códons com informação classificatória,
3. Encontrar a menor subsequência com o maior teor de informação classificatória, resolvendo um problema de otimização com 2 objetivos conflitantes.

Estas 2 últimas tarefas permitem construir conjuntos de dados com tamanho reduzido das sequências codificantes para o estudo da evolução viral mediante a reconstrução de árvores filogenéticas, o que impacta positivamente o custo computacional deste processo habitual em Bioinformática.

7.2 CLASSIFICANDO SEQUÊNCIAS COM O MODELO CBUC

Uma vez que o modelo é treinado e os G agrupamentos são encontrados, estes são anotados por especialista humano, atribuindo a cada agrupamento o tipo/subtipo viral correspondente. O critério de sucesso desta abordagem requer que exista um mapeamento correto entre os agrupamentos achados e os tipos/subtipos virais presentes no dataset de treinamento. Em caso de insucesso, esta abordagem não deve ser utilizada.

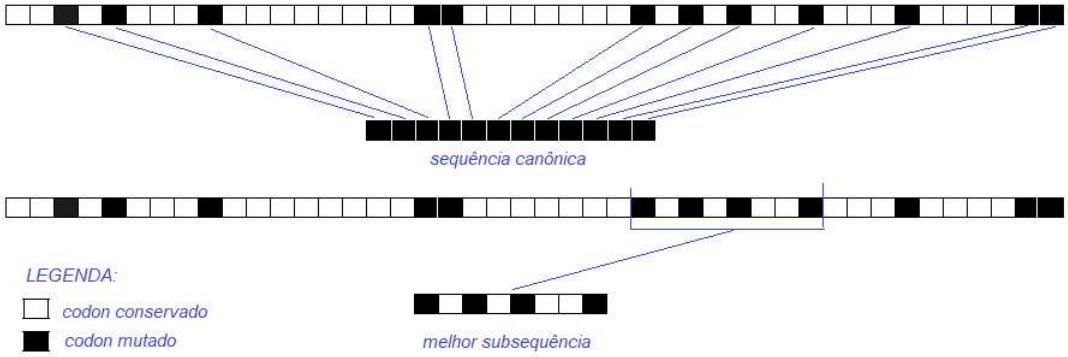
Na hipótese de sucesso, o modelo construído pode ser usado de forma confiável para a classificação de sequências fora da amostra de treinamento. Os pré-requisitos são que a sequência de entrada deve estar no formato FASTA, estar corretamente alinhada e ter o mesmo comprimento que as sequências do conjunto de dados de treinamento. Para ajudar no alinhamento de novas sequências, uma sequência de referência do conjunto de dados de treinamento é selecionada para ser usada como modelo. Um grande número de identidades com a sequência de referência é um bom indicador da qualidade da sequência a ser classificada. Caso contrário, a sequência deve ser desconsiderada, antes mesmo de ser processada pelo classificador. O valor limiar de identidades a ser usado pode ser calibrado a partir dos resultados do modelo, já que $N_c - F$ é uma boa estimativa do número de códons conservados nas sequências estudadas. Desta forma, $(N_c - F)/N_c$ é um limite inferior estatisticamente fundamentado da fração de identidades esperada, se a sequência em estudo for de qualidade.

O processamento de uma nova sequência, corretamente formatada e de boa qualidade, é feito pelo cálculo de sua etiqueta $\mathscr{T}$, que é comparada com as D etiquetas do conjunto de etiquetas T (7.13). Uma vez que cada etiqueta está associada a um único agrupamento e cada agrupamento a um tipo ou subtipo viral, é possível, desta forma, atribuir o tipo/subtipo à sequência que está sendo classificada.

Caso a etiqueta da sequência de entrada não se encontre em T, deve-se fazer uma análise detalhada, reavaliando sua qualidade. Caso a reavaliação da qualidade seja positiva, a sequência deve ser adicionada ao conjunto de treinamento, realizando-se um novo treinamento para atualização do modelo de classificação. Desta forma, um modo de retreinamento sob demanda pode ser programado.

7.3 CONSTRUINDO SEQUÊNCIAS SINTÉTICAS ALTAMENTE INFORMATIVAS

Figura 14 – Sequência Canônica e Melhor Subsequência



Fonte: Imagem do autor.2021

Sequências sintéticas altamente informativas (Figura 14) podem ser facilmente construídas concatenando os códons nas posições $p_j,\ j \in [1, F]$ contidos em $\mathscr{F}$ (Eq. 7.8). Diante a possibilidade de construir este tipo de sequências sintéticas, emergem no momento novas questões de pesquisa:

1. Desde o ponto de vista de classificação (tipagem/subtipagem), quantos dos F atributos são necessários para atingir uma certa acurácia, por exemplo 95 %, com sequências não usadas para o treinamento?. Esta informação permitiria construir sequências sintéticas menores.
2. Desde o ponto de vista das técnicas utilizadas para o estudo da história evolutiva dos vírus, a através da evolução molecular dos genes estudados, é importante saber:
    - Quão diferentes/semelhantes são as árvores filogenéticas construídas com a sequência real e com as sintéticas?
    - Como podem ser caraterizadas as diferenças observadas?
    - Como podem ser explicadas essas diferenças?
    - As árvores construídas com sequências sintéticas preservam o relógio molecular?, ou definem um relógio diferente?, ou o relógio molecular capta apenas parcialmente o histórico?, ou a informação temporal é totalmente perdida no processo?
    - Como podem ser exploradas as diferenças para ampliar o conhecimento sobre os vírus e sobre os métodos filogenéticos?

7.4 ENCONTRANDO O INTERVALO MAIS CURTO E INFORMATIVO

É habitual Varrendo o vetor λ (Eq. 7.6) com janelas deslizantes de diferentes comprimentos em códons, é possível identificar o intervalo mais curto e informativo. Para fazer isso, introduzimos duas medidas dentro da janela, chamadas de Densidade da Informação (I_D) e Índice de Variabilidade (V_I).

Considerando uma janela de tamanho w com seu primeiro códon posicionado no códon c na sequência, a medida I_D em tal posição é calculada como:

$$I_D(c,w) = \frac{\sum_{i=1}^{w} \Big(1 \text{ if } (\lambda_{c+i-1} > 1) \mid 0 \text{ caso contrario}\Big)}{w}, \text{ para } c \in [1, N_c - w + 1] \tag{7.15}$$

onde λ's são os componentes do vetor λ. Como o numerador acima é o número de códons com $\lambda > 1$ dentro de uma janela de tamanho w, ele é sempre menor ou igual a w e, portanto, I_D varia no intervalo unitário $[0,1]$. $I_D = 0$ quando nenhum dos códons, na subsequência definida pela janela, é informativo, ou seja, quando $\lambda = 1$ para todos os códons, enquanto $I_D = 1$ quando todos os códons são informativos, ou seja, quando $\lambda \geq 2$ para todos os códons. Por esta razão, ele mede a densidade (fração) de códons informativos em uma subsequência. Portanto, as subsequências mais informativas em qualquer sequência de codificação têm o maior I_D.

O Índice de Variabilidade para a mesma janela é definido como:

$$V_I(c,w) = \frac{\prod_{i=1}^{w} \lambda_{c+i-1}}{2^{w\ I_D(c,w)}}, \text{ para } c \in [1, N_c - w + 1] \tag{7.16}$$

O numerador na equação acima é o número real de sequências de códons distintas de tamanho w que podem teoricamente ser encontradas no intervalo de códons $[c, c+w-1]$ do conjunto de dados alinhado. O denominador é o número mínimo de subsequências distintas que teoricamente podem ser encontradas no intervalo do códon $[c, c+w-1]$. Portanto, V_I é a razão entre o número real e o número mínimo de subsequências distintas que teoricamente podem ser encontradas no intervalo de códons $[c, c+w-1]$, o que implica $V_I \geq 1$. Portanto, as subsequências mais variáveis em qualquer sequência de codificação têm o maior V_I.

Deve-se notar que subsequências com o mesmo I_D podem ter V_I diferentes.

Foi observado em vários experimentos com conjuntos de dados de teste, que os valores máximos das medidas I_D e V_I deslizando uma janela de qualquer tamanho w ao longo

das sequências $c = 1, 2, \ldots, N_c - w + 1$, vamos denotá-lo como

$$I_D^{max}(w) = \max_c(I_D(c,w))$$

e

$$V_I^{max}(w) \max_c(V_I(c,w))$$

variam de maneira oposta com w, mostrando tendências decrescentes e crescentes, respectivamente. Portanto, o produto das versões em escala de $[0,1]$ das duas medidas, vamos denotá-las como

$$\hat{V}_I^{max}(w) = \frac{V_I^{max}(w) - \min_w(V_I^{max})}{\max_w(V_I^{max}) - \min_w(V_I^{max})}$$

e

$$\hat{I}_D^{max}(w) = \frac{I_D^{max}(w) - \min_w(I_D^{max})}{\max_w(I_D^{max}) - \min_w(I_D^{max})}$$

deve ter pelo menos um máximo para um valor ideal de w_{opt}. Portanto, o problema de encontrar um tamanho de janela ideal pode ser formulado da seguinte forma: Para maximizar $Q(w)$, chamamos de métrica de qualidade, dada por

$$Q(w) = \hat{I}_D^{max}(w)\hat{V}_I^{max}(w) \tag{7.17}$$

em um determinado domínio de w. No caso de $Q(w)$ ter vários máximos locais, o mais alto deve ser escolhido. Na figura 15 ilustramos a dependência do tamanho da janela w das medidas normalizadas, $\hat{I}_D^{max}(w)$ e $\hat{V}_I^{max}(w)$, e seu produto, $Q(w)$, calculado com o conjunto de dados de treinamento do gene envelope do vírus Zika.

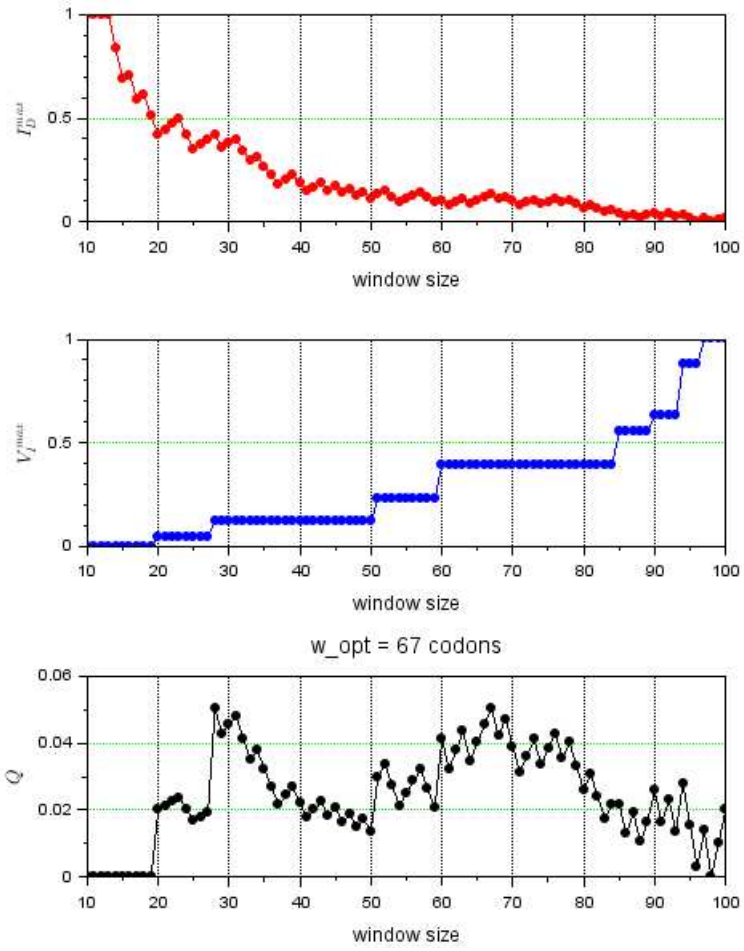
Uma vez encontrado o tamanho ótimo da janela w_{opt}, é necessário encontrar a melhor posição da janela ótima ao longo da sequência, maximizando as duas medidas introduzidas do potencial classificatório de uma sequência de códons. Para fazer isso, avaliamos as medidas I_D e V_I para o tamanho de janela ideal w_{opt} e reescalamos para o intervalo $[0,1]$. Vamos denotá-los como

$$V_I^{opt}(c) = \frac{V_I(c,w_{opt}) - \min_c(V_I(c,w_{opt}))}{\max_c(V_I(c,w_{opt})) - \min_c(V_I(c,w_{opt}))}$$

e

$$I_D^{opt}(c) = \frac{I_D(c,w_{opt}) - \min_c(I_D(c,w_{opt}))}{\max_c(I_D(c,w_{opt})) - \min_c(I_D(c,w_{opt}))}$$

Figura 15 – I_D e V_I reescaladas e seu produto Q em função do tamanho da janela w.



Fonte: Imagem do autor.2021

Como não se espera nenhuma tendência regular das medidas com a posição c da janela ao longo das sequências de códons, construímos uma função objetivo a ser maximizada, $Q^{opt}(c)$, como a soma dos medidas em escala $V_I^{opt}(c)$ e $I_D^{opt}(c)$. Mais especificamente,

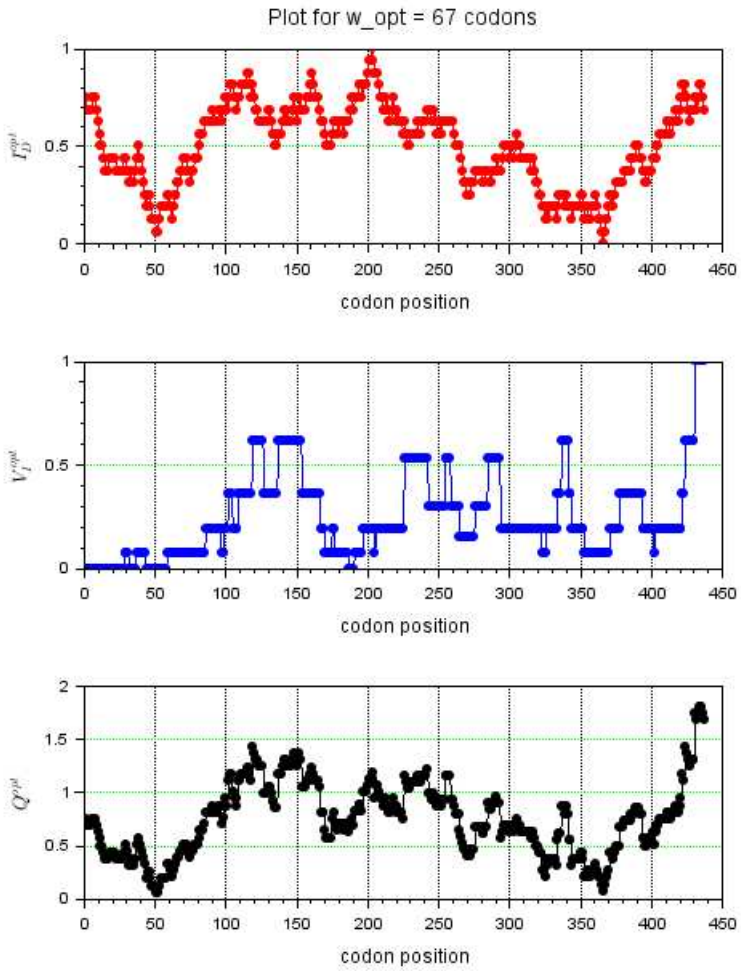
$$Q^{opt}(c) = I_D^{opt}(c) + V_I^{opt}(c), \tag{7.18}$$

Na figura 16 ilustramos a dependência da posição da janela c de $I_D^{opt}(c)$ e $V_I^{opt}(c)$, e sua soma, $Q^{opt}(c)$, calculada com o conjunto de dados de treinamento do gene envelope do vírus Zika.

Procurando o máximo de $Q^{opt}(c)$ encontramos c_{opt} que é a posição do primeiro códon da janela ótima, de tamanho w_{opt}. Em outras palavras, a subsequência do códon começando no códon c_{opt} e se estendendo até o códon $c_{opt} + w_{opt} - 1$, é a subsequência mais curta e mais informativa (Figura 14) .

Deve-se notar que na função objetivo 7.18 usamos medidas reescaladas para dar

Figura 16 – As medidas reescaladas I_D^{opt} e V_I^{opt} e sua soma Q^{opt} em função da posição da janela c.



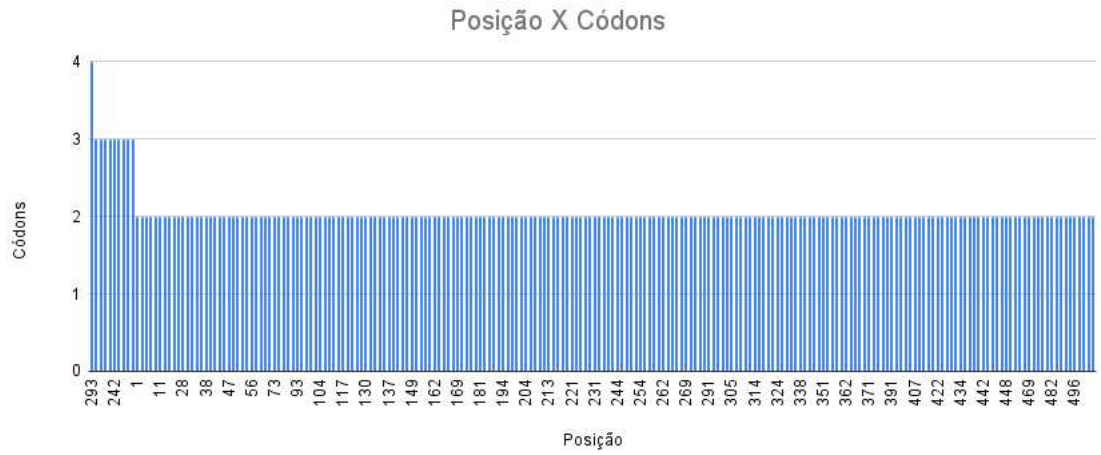
Fonte: Imagem do autor.2021

igual importância à densidade da informação e ao índice de variabilidade ao definir a posição ideal da janela ótima. Pesos diferentes podem ser usados, mas neste caso nosso procedimento para encontrar a subsequência mais curta e mais informativa não é livre de parâmetros. Se apenas uma das medidas for selecionada, deve ser V_I, porque como dito antes, duas sequências com o máximo I_D podem ter V_I diferentes.

8 RESULTADOS

Inicialmente, o método desenvolvido percorreu todo o dataset, agrupando os nucleotídeos encontrados em códons (Seqüência de três nucleotídeos). Como foi utilizado o trecho referente à proteína estrutural, envelope (E), composta por 1512 posições, foram formados 504 grupos de códons em cada seqüência analisada. Então, ao agrupar os códons por posições, identificamos 255 pontos classificatórios, isto é, posições com dois ou mais códons, como mostra a Figura 17.

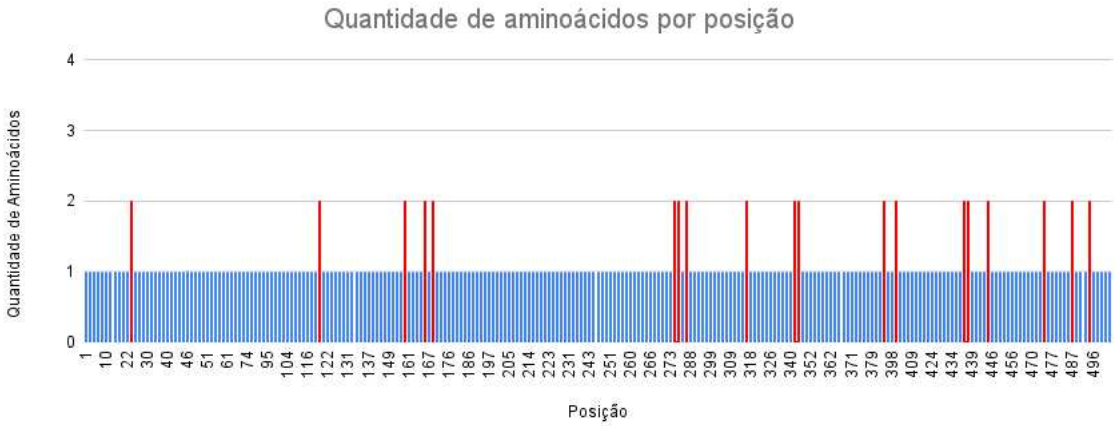
Figura 17 – Posição X Códons



Fonte: Imagem do autor.2021

Conforme foi apresentado na figura 18, temos 255 posições com variedade genética, sendo que, dentre os 61 códons codificáveis, foram encontrados 59 distintos. Mas a variedade identificada nestes pontos, não ocorre quando analisamos os aminoácidos constituídos pelos mesmos, pois o aminoácidos pode ser codificado por até 4 códons. Então, ao buscar a diversidade de aminoácidos em cada posição, encontramos 236 posições com apenas 1 aminoácido e 19 posições com 2 ou mais aminoácidos, processos que são denominados mutação sinônimas e mutação de substituição, respectivamente. Esses dois processos podem ser observados graficamente na Figura 18.

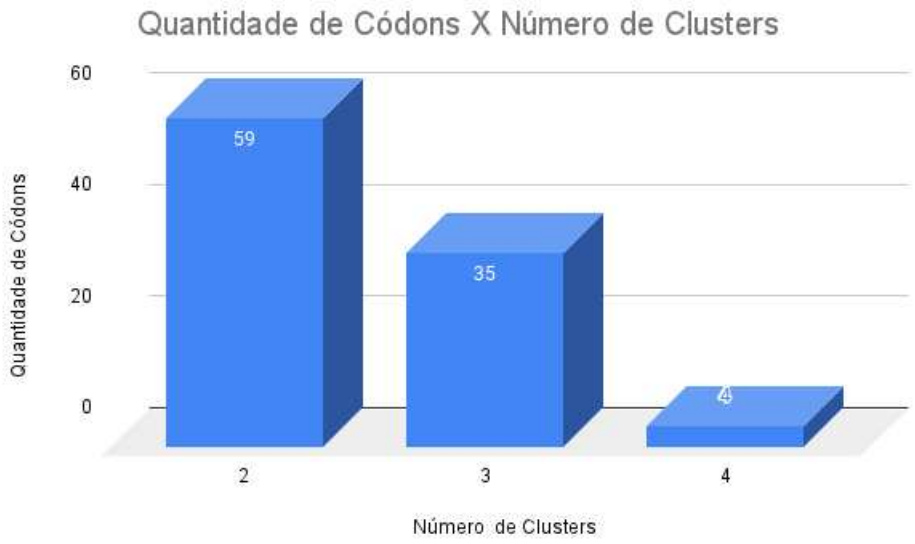
Figura 18 – Quantidade de Aminoácidos por Posição



Fonte: Imagem do autor.2021

Os pontos em vermelho no gráfico são chamados de mutação de substituição, em que, um códon é substituído por outro códon, levando a alteração da matriz de leitura dos aminoácidos e uma possível alteração na proteína que será produzida. Evento que não acontece nos pontos azuis do mapa, chamados de mutações silenciosas, pois as bases encontradas codificam o mesmo aminoácido. Após comprovar a existência de áreas com evidencias de mutações, começamos a construir os clusters de cada posição, que consiste em obter a quantidade de códons diferentes encontradas em cada posição e enumerar-los. Portanto, encontramos o máximo de 4 cluster por posição, como mostra o figura 19. Em um conjunto composto por 61 códons codificantes, encontramos 3 subconjuntos complementares de 2, 3 e 4 clusters, com até 59 códons.

Figura 19 – Quantidade de Códons X Número de Clusters



Fonte: Imagem do autor.2021

Ao utilizar o numero dos clusters para substituir o códon equivalente em cada posição, constituímos um código para cada sequencia utilizada no dataset. Então, ao utilizar técnicas de clusterização com o intuito de ler todos esses códigos e agrupá-los em famílias. Logo, encontramos 3 famílias distintas, sendo que, em cada uma encontramos somente códigos correspondentes a sequencia de um subtipo do vírus da Zika, como mostra a tabela 3.

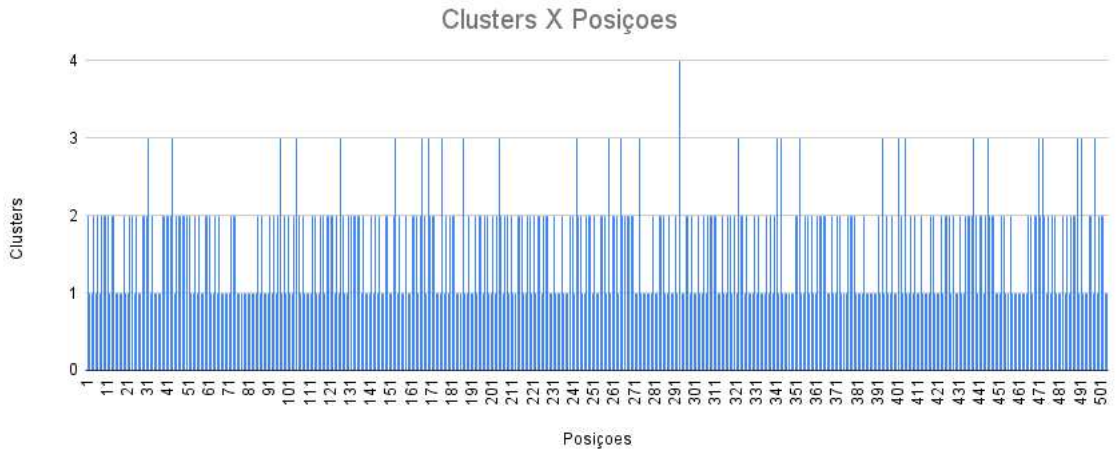
Tabela 3 – Tabela de famílias

Descrição	Accession	SubTipo	Código Sequência
Família 1	KR872956	Asian	2,2,2,1,2,2,1,1,1,1,1,1,2,1,1,1,1,1,2,1,2,2,1,1,1,1,1,2,2,2,1,2,2,1,2,2,2,2,2,2,2,2,2,1,2,2,2,1,1,1,1,1,2,1,1, ...
	KX447511		2,2,2,1,2,2,1,1,1,1,1,1,2,2,1,1,1,1,1,2,1,2,2,1,1,1,1,1,2,2,2,1,2,2,1,2,2,2,2,2,2,2,2,2,1,2,2,2,1,1,1,1,2,1,2,1,1, ...
	KU758877		2,2,2,1,2,2,1,1,1,1,1,1,2,2,1,1,1,1,1,2,1,2,2,1,1,1,1,1,2,2,2,1,2,2,1,2,2,2,2,2,2,2,2,2,1,2,2,2,1,1,1,1,2,1,2,1,1, ...
	KX262887		2,2,2,1,2,2,1,1,1,1,1,1,2,2,1,1,1,1,1,2,1,2,2,1,1,1,1,1,2,2,2,1,2,2,1,2,2,2,2,2,2,2,2,2,1,2,2,2,1,1,1,1,2,1,2,1,1, ...
	MF073358		2,2,2,1,2,2,1,1,1,1,1,1,2,2,1,1,1,1,1,2,1,2,2,1,1,1,1,1,2,2,2,1,2,2,1,2,2,2,2,2,2,2,2,2,1,2,2,2,1,1,1,1,2,1,2,1,1, ...
	KX447515		2,2,2,1,2,2,1,1,1,1,1,1,2,2,1,1,1,1,1,2,1,2,2,1,1,1,1,1,2,2,2,1,2,2,1,2,2,2,2,2,2,2,2,2,1,2,2,2,1,1,1,1,2,1,2,1,1, ...
	MF073357		2,2,2,1,2,2,1,2,1,1,1,2,2,1,1,1,2,1,2,1,2,2,1,1,1,1,1,2,2,2,1,2,2,1,2,2,2,2,2,2,2,2,2,1,2,2,2,1,1,1,1,2,1,2,1,1, ...
	MF073358		2,2,2,1,2,2,1,1,1,1,1,1,2,2,1,1,1,1,1,2,1,2,2,1,1,1,1,1,2,2,2,1,2,2,1,2,2,2,2,2,2,2,2,2,1,2,2,2,1,1,1,1,2,1,2,1,1, ...
	MF073359		2,2,2,1,2,2,1,1,1,1,1,1,2,2,1,1,1,1,1,2,1,2,2,1,1,1,1,1,2,2,2,1,2,2,1,2,2,2,2,2,2,2,2,2,1,2,2,2,1,1,1,1,2,1,2,1,1, ...
	MK238035		2,2,2,1,2,2,1,1,1,1,1,1,2,2,1,1,1,1,1,2,1,2,2,1,1,1,1,1,2,2,2,1,2,2,1,2,2,2,2,2,2,2,2,2,1,2,2,2,1,1,1,1,2,1,2,1,1, ...
	MK566202		2,2,2,1,2,2,1,1,1,1,1,1,2,2,1,1,1,1,1,2,1,2,2,1,1,1,1,1,2,2,2,1,2,2,1,2,2,2,2,2,2,2,2,2,1,2,2,2,1,1,1,1,2,1,2,1,1, ...
Família 2	KU720415	East_African	2,1,2,1,1,2,2,2,2,2,1,2,2,1,2,2,2,2,1,2,1,2,2,2,2,2,2,1,1,1,1,1,2,2,1,1,1,1,2,2,1,1,2,1,2,1,1,2,2,1,2,1, ...
	KY989511		2,1,2,1,1,2,2,2,2,2,2,1,2,2,1,2,2,2,2,2,1,2,1,2,2,2,2,2,2,1,1,1,1,1,2,2,1,1,1,1,2,2,2,1,1,2,1,2,1,1,2,2,1,2,1, ...
	KX377335		2,1,2,1,1,2,2,2,2,2,2,1,2,2,1,2,2,2,2,2,1,2,1,2,2,2,2,2,2,1,1,1,1,1,2,2,1,1,1,1,2,2,2,1,1,2,1,2,1,1,2,2,1,2,1, ...
	KY288905		2,2,2,1,1,2,2,2,2,2,2,2,2,1,1,2,2,2,2,2,1,2,1,2,1,2,1,1,2,1,1,1,2,1,1,2,2,1,1,2,1,1,1,2,1,2,1,1,2, ...
	KY989511		2,1,2,1,1,2,2,2,2,2,1,2,2,1,2,2,2,2,1,2,1,2,2,2,2,2,2,1,1,1,1,1,2,2,1,1,1,1,2,2,2,1,1,2,1,2,1,1,2,2,1,2,1, ...
Família 3	KU955591	West_African	1,2,1,1,1,2,2,2,2,2,1,2,2,2,1,2,2,2,2,1,2,1,1,2,2,2,1,2,1,2,1,1,2,2,2,1,2,1,1,2,2,2,1,1,1,1,1,2,2,1,2,1, ...
	MF510857		2,2,1,1,1,2,2,2,2,2,1,2,2,2,1,2,2,2,2,1,2,1,1,2,2,2,1,2,1,2,1,1,2,2,2,1,2,1,1,2,2,2,1,1,1,1,1,2,2,1,2,1, ...
	KU955592		1,2,1,1,1,2,2,2,2,2,1,2,2,2,1,2,2,2,2,1,2,1,1,2,2,2,1,2,1,2,1,1,2,2,2,1,2,1,1,2,2,2,1,1,1,1,1,2,2,1,2,1, ...
	KU955595		1,2,1,1,1,2,2,2,2,2,1,2,2,2,1,2,2,2,2,1,2,1,1,2,2,2,1,2,1,2,1,1,2,2,2,1,2,1,1,2,2,2,1,1,1,1,1,2,2,1,2,1, ...
	KX198134		1,2,1,1,1,2,2,2,2,2,1,2,2,2,1,2,2,2,2,1,2,1,1,2,2,2,1,2,1,2,1,1,2,2,2,1,2,1,1,2,2,2,1,1,1,1,1,2,2,1,2,1, ...
	KY348860		1,2,1,1,1,2,2,2,2,2,1,2,2,2,1,2,2,2,2,1,2,1,1,2,2,2,1,2,1,2,1,1,2,2,2,1,2,1,1,2,2,2,1,1,1,1,1,2,2,1,2,1, ...
	MF510857		2,2,1,1,1,2,2,2,2,2,1,2,2,2,1,2,2,2,2,1,2,1,1,2,2,2,1,2,1,2,1,1,2,2,2,1,2,1,1,2,2,2,1,1,1,1,1,2,2,1,2,1, ...
	MG758785		1,2,1,1,1,2,2,2,2,2,1,2,2,2,1,2,2,2,2,1,2,1,1,2,2,2,1,2,1,2,1,1,2,2,2,1,2,1,1,2,2,2,1,1,1,1,1,2,2,1,2,1, ...
	MG758786		1,2,1,1,1,2,2,2,2,2,1,2,2,2,1,2,2,2,2,1,2,1,1,2,2,2,1,2,1,2,1,1,2,2,2,1,2,1,1,2,2,2,1,1,1,1,1,2,2,1,2,1, ...
	MK028860		1,2,1,1,1,2,2,2,2,2,1,2,2,2,1,2,2,2,2,1,2,1,1,2,2,2,1,2,1,2,1,1,2,2,2,1,2,1,1,2,2,2,1,1,1,1,1,2,2,1,2,1, ...
	MN025403		2,2,1,2,1,2,2,2,2,2,1,2,2,2,1,2,2,2,2,1,2,1,2,1,1,2,2,2,1,2,1,1,2,1,2,1,1,2,1,2,1,2,2,2,1,1,1,1,2,2,2,1,2,1, ...

Fonte: Imagem do autor.2021

Então, com a intenção de encontrar a região com maior informação filogenética, o CBUC analisa as posições e seus respectivos clusters(Figura 20), recortando em N janelas de N tamanhos, e aplicando as mesmas, as medidas de densidade da informação e índice de variabilidade. Portando, ao utilizar os dados obtidos ao processar o dataset com sequencias do Gene ENV do Zika Vírus, encontramos a região iniciada no códon 434 e finalizada no códon 500.

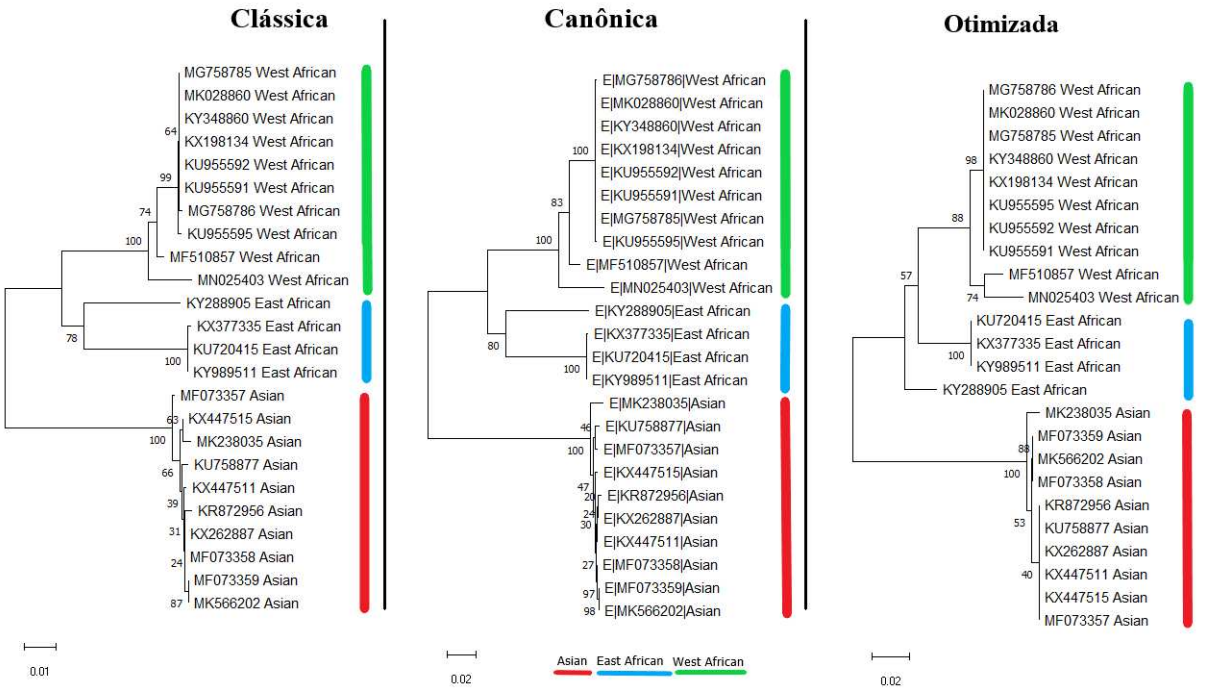
Figura 20 – Posições X Número de Clusters



Fonte: Imagem do autor.2021

Com a finalidade de validar os resultados obtidos pelo CBUC, utilizamos o software MEGA para construir três árvores filogenéticas com os dados resultantes do processamento dos dataset do vírus da Zika e da Dengue. A primeira foi a árvore clássica, composta pelas sequencias completas do dataset. A segunda foi a árvore canônica (Figura 21), árvore filogenética construída com a menor sequência codificante totalmente informativa, composta por toda a informação filogenética extraível de qualquer conjunto de sequências codificantes (Figura 14). E a terceira e última, foi a árvore otimizada , composta pela subsequência que possui o maior número de códons não conservados com o menor comprimento, e codifica o trecho mais variável da cadeia de aminoácidos da proteína codificada pela sequência codificante completa (Figura 21).

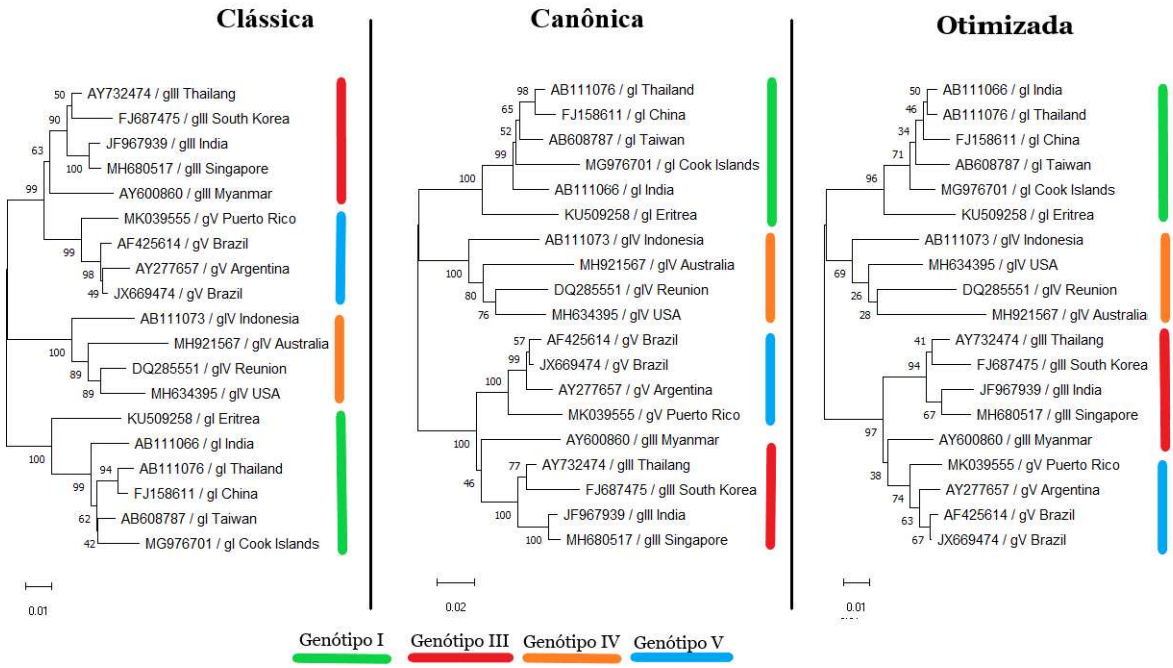
Figura 21 – Árvores Filogenéticas do Zika Vírus



Fonte: Imagem do autor.2021

Desde um ponto de vista teórico, é de esperar que as árvores clássica e canônica (Figura 21) sejam exatamente iguais, levando em conta que os sítios não mutados (conservados) não contribuem no calculo das distâncias. Qualquer diferença observada aponta para alguma inconsistência no modelo de distância utilizado pelo método filogenético usado no nosso estudo. É preciso neste ponto destacar que, como selecionamos os códons mutados, ou seja tripletes de nucleotídeos, e na grande maioria deles, apenas 1 nucleotídeo é mutado, 2 de cada 3 nucleotídeos concatenados na sequência sintética são conservados, o que faz que apenas 1/3 dos nucleotídeos seja informativo desde o ponto de vista de classificação filogenética. Uma versão do método CBUC pode ser implementada a nível de nucleotídeos, que poderia ser chamado método NBUC (*Nucleotideo Based Unsupervised Classification*) , pois permitiria construir sequencias sintéticas somente com nucleotídeos informativos, embora a sequência ótima ainda teria nucleotídeos não informativos.

Figura 22 – Árvores Filogenéticas do Vírus da Dengue



Fonte: Imagem do autor.2021

Ao gerar a arvore filogenética otimizada do vírus da Dengue (Figura 22), notamos que o genótipo gIII AY600860 aparece agrupado ao gV. Este erro é compreensivo, já que a subsequência não carrega toda informação filogenética e esta sequência é a fronteira do genótipo III, tendo com vizinho mais próximo, o genótipo gV, então, é natural que o mesmo seja associado ao genótipo V (Figura 22). Mas este esse erro deixa a seguinte questão para trabalho futuros: Qual é o tamanho ideal de uma subsequência otimizada? Outro ponto observado ao analisar os resultados do vírus da Dengue, foi o número mínimo de sequencias de um mesmo tipo viral para forma um família. O CBUC sempre irá associar sequencias isoladas de um genótipo a algum dos genótipos identificados. Uma sequencia não define família, no mínimo precisam de duas. Logo, não utilizamos o genótipo II do vírus da Dengue, pois não encontramos sequências suficientes do mesmo.

9 CONCLUSÃO

Esta pesquisa trouxe como objetivo geral, a aplicação de um novo método de análise de sequências no estudo de genomas de vírus de importância epidemiológica, visando simplificar o processo de genotipação. Então, foram preparado alguns datasets com sequência de variados tipos de arbovírus e foi desenvolvido o método CBUC(Condon Based Unsupervised Classification) para analisar as sequencias de cada vírus, encontrando regiões codificantes e ao final determinar área com maior informação filogenética.

Conforme mostrado na sessão de resultado desta monografia, ao processar o dataset do vírus da Zika, foram encontrados todos os pontos com códons codificantes, três famílias de padrões que correspondiam ao tipos estudados deste arbovírus e a região com maior informação filogenética.

Com o intuito de validar estes resultados, foram geradas três árvores filogenéticas com o dataset completo, com os pontos que contêm códons codificantes e com a região determinada pelo CBUC, como a área com maior informação filogenética. Os resultados foram positivos para o dataset do Zika Vírus, pois as árvores foram iguais, levando em conta que os sítios não mutados não contribuíram no cálculo das distancias. Mas os resultados do dataset da Dengue apresentaram alguns erros que geraram questões a serem resolvidas em trabalhos futuros.

10 TRABALHOS FUTUROS

Este trabalho abre algumas possibilidades de continuação como citado a seguir:

- Qual é o menor tamanho de uma subsequência ótima, que carrega informação parcial, terá um resultado igual a sequência canônica, que carrega toda a informação?
- Quanta informação filogenética é necessária para obter um agrupamento similar ao agrupamento canônico? Isto pode ser feito determinando o menor número de códons da sequência canônica com os quais se obtém uma arvore similar à canônica. É primordial o estudo em uma gama representativa de vírus.
- A ordenação dos códons da sequência canônica de maior a menor número de clusters é necessário para realizar este último estudo? Assim como introduzir uma medida de quantidade de informação que leve em conta não apenas o número de códons da sequência canônica utilizada, mas também o números clusters de cada um.

REFERÊNCIAS

ALBERTS, B.; JOHNSON, A.; LEWIS, J.; RAFF, M.; ROBERTS, K.; WALTER, P. Biologia molecular da célula. 5ª edição. 2010. **Editora Artmed**.

A¡N, M. A. M.; CAICEDO, M. A.-a. I. A. T.; TORRES, A. K. DÃaz. TÃde BiologÃa Molecular en el desarrollo de la investigaciÃ. RevisiÃde la literatura. **Revista Habanera de Ciencias MÃ**, scielocu, v. 16, p. 796 – 807, 10 2017. ISSN 1729-519X. Disponível em: <http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1729-519X2017000500012&nrm=iso>.

F., M.; P., C. Algorithms for hierarchical clustering: an overview. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 2, n. 1, p. 86–97, 2012.

GERARD, T.; BERDELL, F.; CHRISTINE, C. **Microbiologia. 10ª Edição**. [S.l.]: Editora Artmed. Pg, 2012.

KAISER, T. J. S.; BENICIO, A. A. Alinhamento dos aminoácidos para o sequenciamento genético de proteínas. **JORNADA CIENTÍFICA DA UNESC**, n. 1, 2015.

MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. In: **Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.** [S.l.]: University of California Press, 1967. v. 1, p. 281–297.

MOREL, V.; MARTIN, E.; FRANÇOIS, C.; HELLE, F.; FAUCHER, J.; MOUREZ, T.; CHOUKROUN, G.; DUVERLIE, G.; CASTELAIN, S.; BROCHOT, E. A simple and reliable strategy for bk virus subtyping and subgrouping. **Journal of Clinical Microbiology**, Am Soc Microbiol, p. JCM–01180, 2017.

OLIVEIRA, M. B. S. C. d.; RIBEIRO, F. C. et al. Virologia. In: . [S.l.]: EPSJV, 2009.

PROSDOCIMI, F.; COUTINHO, G.; NINNECW, E.; SILVA, A. F.; REIS, A. N. dos; MARTINS, A. C.; SANTOS, A. C. F. dos; JÚNIOR, A. N.; FILHO, F. C. Bioinformática: manual do usuário. **Biotecnologia Ciência & Desenvolvimento**, v. 29, p. 12–25, 2002.

PURDY, M. A.; SUE, A. The effect of phylogenetic signal reduction on genotyping of hepatitis e viruses of the species orthohepevirus a. **Archives of virology**, Springer, v. 162, n. 3, p. 645–656, 2017.

RITTER, M. d. N.; THEY, N. H.; KONZEN, E. R. Introdução ao software estatístico r. UFRGS. Campus Litoral Norte, 2019.

SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. **Molecular Biology and Evolution**, v. 4, n. 4, p. 406–425, 1987.

TORTORA, G. J.; CASE, C. L.; FUNKE, B. R. **Microbiologia-12ª Edição**. [S.l.]: Artmed Editora, 2016.

VERLI, H. Bioinformática: da biologia à flexibilidade molecular. Sociedade Brasileira de Bioquímica e Biologia Molecular, 2014.

VIANA, G. V. R. Técnicas para construção de árvores filogenéticas. 2007.

WAGGENER, B. (Ed.). **Pulse Code Modulation Techniques**. New York: Springer, 1995. ISBN 9780442014360.

ZAHA, A.; FERREIRA, H. B.; PASSAGLIA, L. M. **Biologia Molecular Básica-5**. [S.l.]: Artmed Editora, 2014.

ZHANG, W.; CHIPMAN, P. R.; CORVER, J.; JOHNSON, P. R.; ZHANG, Y.; MUKHOPADHYAY, S.; BAKER, T. S.; STRAUSS, J. H.; ROSSMANN, M. G.; KUHN, R. J. Visualization of membrane protein domains by cryo-electron microscopy of dengue virus. **Nature Structural & Molecular Biology**, Nature Publishing Group, v. 10, n. 11, p. 907–912, 2003.