RUITHER ARTHUR LOCH GOMES

CLASSIFICAÇÃO E ANOTAÇÃO *in silico* DE GENOMAS VIRAIS RELACIONADOS AO FILO *Cressdnaviricota*

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Bioquímica Aplicada, para obtenção do título de *Doctor Scientiae*

Orientador: Francisco Murilo Zerbini Junior

VIÇOSA - MINAS GERAIS
2023

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
G633c
2023

Gomes, Ruither Arthur Loch, 1990-
Classificação e anotação *in silico* de genomas virais relacionados ao filo *Cressdnaviricota*: / Ruither Arthur Loch Gomes. – Viçosa, MG, 2023.
1 tese eletrônica (79 f.): il. (algumas color.).

Texto em português e inglês.
Orientador: Francisco Murilo Zerbini Júnior.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Fitopatologia, 2023.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2024.002
Modo de acesso: World Wide Web.

1. Vírus - Genética. 2. Mapeamento cromossômico. 3. Aprendizado do computador. 4. Redes neurais (Computação). I. Zerbini Júnior, Francisco Murilo, 1966-. II. Universidade Federal de Viçosa. Departamento de Fitopatologia. Programa de Pós-Graduação em Bioquímica Aplicada. III. Título.

CDD 22. ed. 579.24

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

RUITHER ARTHUR LOCH GOMES

CLASSIFICAÇÃO E ANOTAÇÃO *in silico* DE GENOMAS VIRAIS RELACIONADOS AO FILO *Cressdnaviricota*

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Bioquímica Aplicada, para obtenção do título de *Doctor Scientiae*.

APROVADA: 02 de maio de 2023

Assentimento:

gov.br Documento assinado digitalmente
RUITHER ARTHUR LOCH GOMES
Data: 29/01/2024 18:11:14-0300
Verifique em https://validar.iti.gov.br

Ruither Arthur Loch Gomes
Autor

gov.br Documento assinado digitalmente
FRANCISCO MURILO ZERBINI JUNIOR
Data: 29/01/2024 18:03:52-0300
Verifique em https://validar.iti.gov.br

Francisco Murilo Zerbini Junior
Orientador

AGRADECIMENTOS

Aos meus pais, Ivete e Luiz Pereira, que sempre me incentivaram e apoiaram de todas as formas que podem, a sempre estudar, me importar e seguir minhas paixões.

Às minhas irmas Rayana e e Rayara pelo carinho e amor que sempre tivemos e teremos um com os outros .

À minha esposa Rayane, que com seu companheirismo, amor, paciencia e esforço, mudou a minha vida e me faz querer sempre buscar o melhor, para retribuir tanta sorte de encontrar alguém assim.

A meus "filhos" de quatro patas, que com seu amor e brincadeiras, me tranquilizam e ajudam a enfrentar os desafios.

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Genética e Melhoramento pela oportunidade em realizar o doutorado e conduzir esta pesquisa.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão das bolsas de estudos.

Ao professor Francisco Murilo Zerbini Júnior que foi essencial para eu chegar até aqui, pelos conselhos e amizade, nos incentivando a polir o senso critico, por toda dedicação e paciência.

Aos amigos de laboratório, Tarsiane, Ayane, Lucas, Roberta, João Paulo, Osvaldo, Baltazar, Patrícia, pelo companheirismo, ajudas, conversas e momentos de descontração.

A todos que de alguma forma contribuíram para realização deste trabalho:

Muito Obrigado!

RESUMO

GOMES, Ruither Arthur Loch, D.Sc., Universidade Federal de Viçosa, maio de 2023. **Classificação e anotação *in silico* de genomas virais relacionados ao filo *Cressdnaviricota*.** Orientador: Francisco Murilo Zerbini Júnior.

Os vírus afetam ciclos biogeoquímicos e infectam organismos em todos os ambientes da terra. Avanços em diferentes tecnologias, como o sequenciamento de alto rendimento e a biologica computacional, trouxeram luz sobre a real diversidade e abundância dos vírus. Uma das consequências mais importantes foi a descoberta de um imenso número de sequências virais, porém sem similaridade com vírus previamente caracterizados. Enquanto a classificação taxonômica dos vírus havia sido feita por décadas com base em características fenotípicas, essa nova realidade gerou a necessidade da utilização direta das sequências, mesmo na ausência de qualquer informação biológica, para a classificação taxonômica. Com esse novo panorama de farta disponibilidade de dados de sequência, avanços no poder computacional e de aprendizado de máquina surgiram como ferramentas essenciais para classificação e anotação das sequências derivadas desse "dilúvio de dados". Diversas ferramentas computacionais vêm sendo propostas e desenvolvidas usando diferentes abordagens para trabalhar com esses dados, e o aprendizado de máquina vem se destacando por sua alta acurácia de predição. Na taxonomia, diferentes abordagens vem sendo aplicadas para grupos específicos de vírus, e só recentemente foi desenvolvido um algoritmo, VirusTaxo, para classificação taxonômica de todos os tipos de vírus com acurácia considerável. Entre as diversas famílias de vírus, algumas se enquadram dentro de um grupo de vírus com genomas de DNA de fita simples circulares e pequenos, que codificam uma proteína relacionada à replicação que é relativamente conservada entre seus membros. Esses vírus, classificados no filo *Cressdnaviricota*, são exemplares interessantes para se avaliar métodos *in silico* de classificação e análise de funções gênicas. Assim, na primeira parte desse trabalho, foi avaliada a capacidade das redes neurais convolucionais para classificar taxonomicamente os cressdnavírus. Foi possível obter uma acurácia nos dados de teste superior ao VirusTaxo, a ferramenta com maior capacidade de predição taxonômica atualmente. Na segunda parte, foram utilizadas ferramentas computacionais para identificar possíveis pequenas ORFs funcionais em alfassatélites associados a begomovírus que possam estar relacionadas a variações de sintomas observadas entre alfassatélites do Novo Mundo e do Velho Mundo e foi possível identificar duas pequenas ORF com domínios funcionais preditos.

Palavras-chave: Vírus, *Cressdnaviricota*, bioinformática, machine learning

ABSTRACT

GOMES, Ruither Arthur Loch, D.Sc., Universidade Federal de Viçosa, May 2023. **Classification and *in silico* annotation of viral genomes related to the phylum *Cressdnaviricota*.** Advisor: Francisco Murilo Zerbini Júnior.

Viruses affect biogeochemical cycles and infect organisms in all environments on earth. Advances in different technologies, such as high-throughput sequencing and computational biology, have shed light on the real diversity and abundance of viruses. One of the most important consequences was the discovery of an immense number of viral sequences with no similarity with previously characterized viruses. While the taxonomic classification of viruses had been carried out for decades based on phenotypic characteristics, this new reality created the need for the direct use of sequences, even in the absence of any biological information, for taxonomic classification. With this new scenario of plentiful availability of sequence data, advances in computational power and machine learning have emerged as essential tools for classifying and annotating sequences derived from this "data deluge". Several computational tools have been proposed and developed using different approaches to work with these data, and machine learning has been highlighted for its high prediction accuracy. In taxonomy, different approaches have been applied to specific groups of viruses, and only recently an algorithm, VirusTaxo, has been developed to taxonomically classify all types of viruses with considerable accuracy. Among the many families of viruses, some fall within a group of viruses with small, circular, single-stranded DNA genomes that encode a replication-related protein that is relatively conserved among its members. These viruses, classified in the phylum *Cressdnaviricota*, are interesting examples for evaluating *in silico* methods of classification and analysis of gene function. Thus, in the first part of this work, the ability of convolutional neural networks to taxonomically classify cressdnaviricots was evaluated. It was possible to obtain an accuracy in the test data superior to VirusTaxo, the tool currently with greater taxonomic prediction capacity. In the second part, computational tools were used to identify possible small functional ORFs in begomovirus-associated alphasatellites that may be related to symptom variations observed between New World and Old World alphasatellites and it was possible to identify two small ORFs with predicted functional domains.

Keywords: Virus, *Cressdnaviricota*, bioinformatics, machine learning

SUMÁRIO

INTRODUÇÃO GERAL

Desde sua origem no final do século 19, a virologia gerou descobertas que nos surpreendem e reforçam o pouco conhecimento que temos sobre os vírus. Ao longo do século 20, seu impacto na saúde humana, animal e vegetal tornou-se evidente causando pandemias que marcaram a história como a "gripe espanhola" em 1918 e a AIDS na década de 1980 (Hahn *et al.*, 2000). Mais recentemente, a metagenômica revelou a real abundância, diversidade e ubiquidade dos vírus, identificados em todos os ambientes da terra (Gorbalenya *et al.*, 2020; Simmonds *et al.*, 2017). Os vírus afetam biomas e influenciam ciclos biogeoquímicos (Ibrahim *et al.*, 2018). Desse modo, esses antigos (porém só recentemente estudados) membros da biosfera requerem estudos e vigilância constantes.

Inicialmente, a identificação dos vírus foi realizada utilizando-se técnicas de cultivo de células, obtendo-se partículas virais por ultracentrifugação. Apesar de suas limitações essas técnicas foram extremamente úteis na descoberta e tratamento de muitas doenças virais. Assim, até o início da segunda metade do século 20, acreditava-se que a maioria das doenças virais já haviam sido identificadas (e os vírus classificados). Entretanto, fatores como o desmatamento e a perda de habitats, a rápida urbanização muitas vezes mal planejada, e a facilitação de viagens aéreas globais, levam atualmente diferentes populações humanas à convivência e/ou exposição a diferentes animais selvagens (e seus parasitas, como os vírus), facilitando o surgimento de novas doenças (Datta *et al.*, 2015). Apenas no século 21, sete eventos desse tipo, denominados *spillover* zoonótico, afetaram gravemente a população humana: (*i*) SARS-CoV em 2003 na China, (*ii*) influenza H1N1 em 2009 na América do Norte, (*iii*) MERS-CoV em 2012 na Arábia Saudita, (*iv*) ebola em 2014 na Libéria e Serra Leoa, (*v*) zika em 2015 no Brasil, (*vi*) SARS-CoV-2 em 2020 em todo o mundo, e (*vii*) mpox em 2022 em vários países da Europa e das Américas (Tajudeen *et al.*, 2022).

Com o avanço da ciência, técnicas baseadas em sequências de ácidos nucleicos foram desenvolvidas, tornando obsoletas para determinados objetivos as técnicas dependentes de cultivos de células. Técnicas como a hibridização de ácidos nucleicos e o sequenciamento de fragmentos do ácido nucleico viral amplificados via PCR eram muito mais rápidas em comparação com cultura de células, e possibilitaram a descoberta de vários novos vírus. Entretanto, essas técnicas também possuíam limitações, principalmente a dependência de informação prévia da sequência-alvo para o desenho de oligonucleotídeos e de sondas de hibridização. Devido a essas limitações, os novos vírus que foram descobertos com a utilização dessas técnicas eram, em sua absoluta maioria, relacionados aos já conhecidos (Datta *et al.*, 2015).

A descoberta de novos vírus sem relacionamento direto com aqueles já caracterizados somente ocorreu com o desenvolvimento de novas tecnologias de sequenciamento de alto desempenho (*high throughput sequencing*, HTS) (Martinez & Nelson, 2010), que tornaram possíveis os estudos de metagenômica (sequenciamento dos ácidos nucleicos totais presentes em amostras ambientais, como água, solo, ou fezes de animais) (Mizuno *et al.*, 2013; Roossinck *et al.*, 2015; Rosario & Breitbart, 2011; Shi *et al.*, 2016). As tecnologias de sequenciamento utilizadas nos estudos de metagenômica não requerem nenhum tipo de conhecimento prévio sobre o material genético dos organismos presentes na amostra (Datta *et al.*, 2015).

Buscando a descoberta de novos vírus, desenvolveram-se técnicas específicas de amplificação de material genético para metagenômica viral ("metavirômica"), como a amplificação de primer-único independente de sequência (*sequence-independent single-primer amplification*, SISPA), e a amplificação por círculo rolante (*rolling-circle amplification*, RCA), que aumentam especificamente a quantidade de material genético viral presente na amostra, e assim levaram à descoberta de vários novos vírus (Bexfield & Kellam, 2011; Maclot *et al.*, 2020).

Uma das principais descobertas decorrente da metavirômica foi a revelação da imensa diversidade e quantidade de vírus em todos os habitats. Estima-se que se todas as partículas virais fossem enfileiradas de ponta a ponta, percorreriam uma distância maior que a que separa as 60 galáxias mais próximas da Via Láctea (Suttle, 2005). Um único trabalho (Paez-Espino *et al.*, 2016) descreveu mais de 125 mil novos genomas virais, levando a um aumento de dezesseis vezes no número de genes virais conhecidos. Os estudos mais recentes de metagenômica geraram sequências em escala de petabytes (1 PB sendo equivalente a 10^9 MB), e relataram milhões de possíveis novos vírus (Edgar *et al.*, 2022; Neri *et al.*, 2022).

A obtenção de um número elevado de sequências virais sem qualquer informação biológica relacionada, e sem nenhuma similaridade com sequências já conhecidas, consiste em um enorme desafio para a classifição taxonômica, que é realizada pelo Comitê Internacional de Taxonomia de Vírus (*International Committee on Taxonomy of Viruses*, ICTV, https://ictv.global/).

Em suas origens, a taxonomia viral agrupava os vírus com base em propriedades biológicas, como o formato e dimensões da partícula viral, a gama de hospedeiros e a forma de transmissão. Dessa forma, apenas vírus que haviam sido isolados e cultivados (em sua quase totalidade, vírus que impactavam a sociedade) eram classificados (Murphy *et al.*, 1995). Entretanto, com centenas de milhares de novas sequências virais sem similaridade com grupos taxonômicos conhecidos sendo relatadas nos trabalhos de metavirômica, o ICTV passou a aceitar que a classificação pudesse ser feita exclusivamente com base em sequências, sem a necessidade de nenhum tipo de informação biológica (Simmonds *et al.*, 2017). Com a incorporação dessas sequências ao arcabouço taxonômico, foi possível expandir a taxonomia viral, até então restrita às categorias de espécie, gênero, família e ordem, para incluir também as categorias superiores de classe, filo, reino e domínio (Gorbalenya *et al.*, 2020; Koonin *et al.*, 2020).

Os genomas virais carregam informações em diferentes níveis que podem ser úteis para classificação taxonômica, como o tipo e a organização do genoma, a estratégia usada para replicação, proteínas codificadas, as relações evolutivas, composições locais e globais do genoma, entre outras (Simmonds *et al.*, 2023). Entretanto, a falta de ferramentas com precisão e confiabilidade para extrair e inferir todas as informações existentes nas sequências genéticas dos mais diversos grupos virais limita a identificação de muitas das características codificadas. Além disso, existem dificuldades relacionadas às técnicas de metavirômica envolvidas na obtenção da maioria das sequências virais atualmente, como a montagem de genomas artificiais, a distinção entre sequências episomais e vírus integrados nos genomas de hospedeiros, ou a classificação correta de genomas de virus multissegmentados. Assim, embora o ICTV aceite que vírus cujas sequências tenham sido geradas em estudos de metagenômica possam ser classificados, essas sequências devem satisfazer critérios rigorosos de qualidade, tanto em termos de cobertura (devem corresponder no mínimo à região codificadora completa) como de profundidade (para minimizar a taxa de erro) (Simmonds *et al.*, 2023).

Apesar de ainda não existirem ferramentas completas para se realizar todos os diferentes estudos possíveis, vários esforços têm sido feitos para desenvolver novos algoritmos que sejam cada vez mais confiáveis e fáceis de serem implementados. A bioinformática é uma área em rápida evolução, com criação tanto de algoritmos pontuais quanto de plataformas contendo varios algoritmos diferentes, que auxiliam na genômica comparativa, nos estudos de redes de interações, de filogenética e evolução, nas predições de estruturas tridimensionais de RNA e proteínas, nas análises de vias reguladoras e de expressão genética, e outras (Auslander *et al.*, 2021).

Nesse contexto, Pappas *et al.* (2021) reforçam que para estudar a imensa biodiversidade dos vírus, computadores e algoritmos especializados são necessários. Os autores afirmam ainda que, apesar de sempre ser necessário checar na bancada os resultados obtidos de predições

computacionais, recursos financeiros e um tempo precioso podem ser economizados com algoritmos cada vez mais confiáveis. Eles também apontam o enorme potencial do aprendizado de máquina na virologia, permitindo obter predições confiáveis lidando com dados cada vez mais complexos. Finalmente, concluem que a bioinformática já se tornou um componente tão crucial na pesquisa com vírus quanto as técnicas clássicas.

Diferente da taxonomia de organismos celulares, atualmente baseada na filogenia de genes "universais" (principalmente genes que codificam componentes da maquinaria de tradução protéica), nos vírus não existe nenhum gene que seja universalmente conservado (Dutilh *et al.*, 2021; Koonin *et al.*, 2020; Simmonds *et al.*, 2023). Em consequência, a classificação separa os vírus em domínios (*realms*) cujos membros possuem um ou mais genes/características em comum. Para vírus cujos genomas possuem mais de 10.000 nucleotídeos (vírus com genoma de DNA classificados nos domínios *Duplodnaviria* e *Varidnaviria*), o maior número de genes torna possível a classificações baseadas em redes de similaridade das proteínas conservadas entre os grupos. Já para vírus com genomas menores, a classificação é frequentemente baseada em um único gene principal conservado, como no domínio *Riboviria* com a RNA polimerase dependente de RNA (*RNA-dependent RNA polymerases*, RdRp) (Koonin *et al.*, 2020), e no filo *Cressdnaviricota* (domínio *Monodnaviria*), no qual várias ordens e famílias são agrupadas pela filogenia da proteína Rep, envolvida na replicação por circulo rolante (Krupovic *et al.*, 2020).

Desse modo, os trabalhos de desenvolvimento de ferramentas computacionais para a classificação de vírus normalmente focam em abordagens específicas. Uma abordagem frequentemente utilizada é avaliar a organização do genoma quanto ao conteúdo gênico (presença e ordenação de genes homólogos) de forma hierárquica (de acordo com diferentes níveis taxonômicos). Isso foi feito informalmente por muito tempo, mas recentemente surgiram ferramentas como VConTACT (Bolduc *et al.*, 2017), que avalia redes de proteínas conservadas

para vírus de dsDNA que infectam procariotos (aos níveis de gênero a ordem), e GRAVITy (Aiewsakun & Simmonds, 2018), que avalia tanto o conteúdo gênico quanto padrões genômicos característicos para vírus que infectam eucariotos (de gênero a família).

Ainda se baseando em proteínas conservadas dentro de grupos, outra metodologia comumente aplicada é a criação de árvores filogenéticas bayesianas ou de máxima verossimilhança (Guindon *et al.*, 2010; Price *et al.*, 2010). Dependendo dos grupos de vírus, essas análises são feitas a partir de alinhamentos múltiplos de uma única proteína conservada (ou de algumas proteínas de forma concatenada). Entretanto, os algoritmos atualmente existentes são aplicáveis somente para delimitar diferentes vírus ao nível de espécie (Gorbalenya & Lauber, 2022).

Uma outra forma de classificar genomas dentro de grupos taxonômicos virais, aplicável quando a evolução do grupo segue um modelo de relógio molecular conhecido (com taxas de evolução molecular relativamente constantes), é utilizar a divergência genética com base em comparações de sequências "par a par" (porcentagem de nucleotídeos/amino ácidos divergentes entre duas sequências alinhadas), e adotar valores de demarcação dos taxa baseados na distribuição desses valores (adotando valores localizados nos "vales" entre "picos" no gráfico de distribuição). Essa metodologia foi inicialmente aplicada para classificar algumas famílias virais, e mais recentemente foi disponibilizada na forma de uma ferramenta denominada PASC (Bao *et al.*, 2012) (https://www.ncbi.nlm.nih.gov/sutils/pasc/viridty.cgi).

Tanto o PASC quanto outro algoritmo muito usado para essas análises, o SDT (Muhire *et al.*, 2014), possuem uma limitação quanto ao tamanho e presença de viés no conjunto de dados, visto que conjuntos de dados com poucas sequências geram poucas comparações, reduzindo a confiabilidade estatística das frequências de distribuição de valores (Gorbalenya & Lauber, 2022). Uma outra ferramenta que se baseia em alinhamentos par a par de proteínas conservadas e que lida melhor com desbalanceamento dos dados (por utilizar aprendizado de

máquina para identificar os limiares que categorizam diferentes níveis taxonômicos) é o DEmARC (Lauber & Gorbalenya, 2012). Algumas outras ferramentas também auxiliam no agrupamento de diferentes grupos para alguns casos mais específicos, permitindo lidar com grupos que não sejam monofiléticos (Meier-Kolthoff & Göker, 2017; Moraru *et al.*, 2020).

Por fim, existem algumas abordagens que utilizam características mais simples como conteúdo GC ou a contagem de k-mers (total de sequências de *k* nucleotídeos existentes no genoma). Entretanto, mesmo sendo teoricamente mais práticas e fáceis de serem implementadas, essas abordagens são limitadas em relação às técnicas citadas anteriormente, por tenderem a ser mais sensíveis a amostragem e a eventos moleculares como recombinações e variação nas taxas de evolução de regiões específicas do genoma (Gorbalenya & Lauber, 2022).

Embora inicialmente as metodologias para o desenvolvimento de algoritmos tenham sido predominantemente baseadas em alinhamentos de sequências, as técnicas de aprendizado de máquina (*machine learning*, ML) e especialmente aprendizado profundo (*deep learning*, DL) estão sendo cada vez mais comumente incorporadas à bioinformática, com considerável eficácia. Essas técnicas permitem a realização de forma automática das etapas de extração e seleção de características de dados biológicos, tanto com dados rotulados ou não, que são então usados em modelos para diferentes tipos de predições, com valores elevados de acurácia (Auslander *et al.*, 2021).

Especificamente na anotação e classificação dos genomas virais, algumas ferramentas que não usam alinhamentos surgiram recentemente tanto utilizando metodologias simples de k-mers, quanto usando em muitos casos DL, com bons resultados em diferentes tipos de análises, como predição de promotores e outros elementos regulatórios, de RNAs longos não codificantes, de sequências-alvo de miRNAs, e de arranjos CRISPR (Auslander *et al.*, 2021).

Raju *et al.* (2022) publicaram um novo algoritmo denominado VirusTaxo, que é inovador ao ser o primeiro a classificar taxonomicamente tanto vírus de DNA quanto RNA dentro de centenas de gêneros diferentes. Somente gêneros com sequências únicas (singletons) foram removidos do conjunto de dados, compostos de genomas da lista de espécies classificadas pelo ICTV em 2019. O algoritmo obteve uma acurácia média para os gêneros de 93% durante seu treinamento, e identificou um número de contigs de dados de metagenômica equivalente aos identificados utilizando DeepVirFinder (Ren *et al.*, 2020), que usa redes neurais convolucionais (*convolutional neural networks*, CNN), mas não classifica taxonomicamente as sequências identificadas.

Baseadas na percepção visual, CNNs foram criadas a partir das redes neurais artificiais (*artificial neural networks*, ANN) buscando extrair automaticamente dos dados as características que serão usadas no treinamento, e mais recentemente vem sendo usadas com muito sucesso na análise de dados biológicos. As CNNs têm sido usadas principalmente no reconhecimento de imagens usando redes 2D (bidimensionais), que utilizam matrizes como entrada, mas também na análise de dados textuais usando redes 1D (unidimensionais), que utilizam vetores como entrada (Li *et al.*, 2022; Masko & Hensman, 2015). Essas redes são formadas por diferentes camadas, que podem ser agrupadas de acordo com a necessidade.

As camadas de convolução são as principais camadas de uma CNN, onde cada camada convolucional é composta por um número específico (ou seja, definido pelo criador da rede) de filtros, também chamados kernels, que em redes 2D de classificação supervisionada são matrizes com tamanho também específico. De maneira similar ao que ocorre nos "neurônios" de outros tipos de redes de DL, cada kernel é inicializado com valores aleatórios (conhecidos como pesos, sendo estes parâmetros treináveis na rede) em cada posição da matriz e se move pelas duas dimensões da matriz de entrada. A cada passo/posição (*stride*) que o kernel se move pela matriz de entrada (uma direção por vez) é realizado o produto escalar da matriz de entrada

com essa matriz de cada kernel, e ao resultado desse produto é adicionado um valor (também sendo esse um parâmetro treinável) conhecido como viés (*bias*), sendo finalmente esse valor final passado para uma função de ativação não linear, que busca reduzir a linearidade entre entrada e saída. Após serem calculados os valores para cada passo, no final se obtém uma matriz de resultados (também conhecida como mapa de características, *feature map*) como saída para cada kernel, sendo essas matrizes usadas na próxima camada da rede (Li *et al.*, 2022; Masko & Hensman, 2015).

O *padding* é outro parâmetro inicial importante a ser definido em uma camada convolucional. Quando o *padding* é definido como *same*, são adicionados vetores de zeros (o que não ocorre usando a outra opção, *valid*) nos cantos da matriz de entrada (podendo ser mais do que um vetor por canto, dependendo do *stride*), de modo que os valores presentes na borda da matriz inicial sejam levados em consideração mais de uma vez. Isso ocorre porque algumas combinações de tamanhos de matriz e *stride* podem fazer com que os valores das bordas sejam menos usados para obter a matriz de resposta, afetando essa saída (Islam *et al.*, 2021).

Um outro tipo de camada usado em CNNs são as camadas de *pooling*, que reduzem as dimensões de cada matriz de entrada, como as recebidas das camadas de convolução, permitindo a redução no uso dos recursos computacionais com preservação de características de interesse. Isso é realizado definindo o tamanho de uma janela que percorrerá a matriz de entrada como feito na convolução, mas agora usando uma função para obter a resposta, como a média dos valores ou o valor máximo dentro dessa janela (Lecun *et al.*, 2015). Por fim, existem as camadas de *dropout*, que retiram uma porcentagem (onde 0.3 significa 30%) das conexões de neurônios/dados de entrada de forma aleatória. Isso é feito buscando evitar que o modelo memorize os resultados do conjunto de dados de entrada (algo conhecido como *overfitting*) em vez de buscar uma regra preditiva generalizada, não funcionando bem para dados novos (Dietterich, 1995; Srivastava *et al.*, 2014).

Após o uso dessas diferentes camadas em grupos e ordens específicas, a parte convolucional (de extração de dados) da rede é finalizada com uma camada de "achatamento" (*flattening*), que irá converter todas as matrizes recebidas em um único vetor concatenado, que é então dado de entrada para próxima etapa, onde ocorre a predição, usando um número específico de camadas densas das ANNs. Cada camada densa é composta por um número específico de neurônios (de maneira similar aos kernels das camadas convolucionais), de forma que cada neurônio recebe como entrada todos os dados/saídas dos neurônios da camada anterior, multiplicando-se os mesmos por pesos (outro parâmetro treinável inicializado com valor aleatório) diferentes para cada entrada, e somando-se os resultados adicionando também um viés, para então repassar o valor de resposta para uma função de ativação que altera o valor de forma não linear, simulando a ativação de neurônios biológicos (Li *et al.*, 2022; Masko & Hensman, 2015).

Na ultima camada, em casos de problemas de classificação, é usado um número de neurônios igual ao número de categorias sendo classificadas. A função de ativação utilizada, denominada *softmax*, libera um vetor de probabilidades da amostra pertencer a cada categoria, com a soma das probabilidades sendo sempre igual a 1. Assim, durante o treinamento, a diferença entre o valor correto e o predito é computado pela função de custo/perda, e o erro é propagado de volta pela rede, alterando todos os pesos e vieses, buscando determinar aqueles que tornem os resultados dos treinamentos seguintes mais próximos ao resultado esperado (Li *et al.*, 2022; Masko & Hensman, 2015).

Após usar todos os dados de treinamento, a rede prediz os dados teste, obtendo os resultados de desempenho do modelo sobre esses dados, e reinicia um novo treinamento (conhecido como época), buscando novamente atualizar os parâmetros treináveis para melhorar o desempenho do modelo sobre o teste. Como o treinamento pode levar varias épocas até que se obtenha o deempenho desejável, uma ferramenta útil é o *early stopping*, que detecta quando

o modelo apresenta *overfitting* durante o treinamento por meio do monitoramento de alguma métrica, como a acurácia ou a perda nos testes de validação. Desse modo, quando essa métrica atinge um valor definido, o treinamento é interrompido e recupera os parâmetros treináveis da época com métrica ótima (Prechelt, 2012).

Dentre as métricas de desempenho, a àrea sobre a curva ROC (*receiver operating characteristic*) é uma das mais usadas para conjuntos de dados balanceados (Melo, 2013). Já quando lidando com dados não balanceados, a área sobre a curva de *precision-recall* é tida como uma métrica melhor, visto que leva em consideração a precisão do modelo (Sofaer *et al.*, 2019). Outra métrica útil em problemas de classificação é o *F1 score*, que consiste na média harmônica entre precisão e *recall* e é mais afetado pela classificação em si do que pelas probabilidades obtidas dos resultados (Wardhani *et al.*, 2019).

Durante a avaliação do desempenho de diferentes modelos de ML, a separação dos dados em um conjunto para treinamento e outro para testes é frequentemente aplicada para evitar o *overfitting*, treinando o modelo em parte dos dados e avaliando o desempenho em dados ainda não "vistos", confirmando que o modelo não esta somente memorizando as respostas. Isso é conhecido como validação cruzada (*cross-validation*, CV). Um método comumente usado é a CV *K*-Fold, na qual o conjunto de dados é dividido em *K* conjuntos menores. Cada conjunto é usado uma vez como conjunto de teste, enquanto o restante é usado para o treinamento. O processo é repetido *K* vezes, uma vez para cada conjunto menor (Hastie *et al.*, 2013).

Com diversos parâmetros diferentes treináveis e adaptáveis, além de novas aplicações e técnicas surgindo com frequência, as CNNs são como uma metodologia com alto potencial para utilização na classificação taxonômica de vírus. Diversos estudos recentes de metagenômica expuseram o viés de conhecimento relacionado aos vírus com genoma de DNA de fita simples (ssDNA), que por muito tempo foram vistos como um grupo relativamente raro de vírus

(Rosario, K. *et al.*, 2012; Tisza *et al.*, 2020). Atualmente, vírus de ssDNA já foram identificados na maioria dos habitats e infectando hospedeiros diversos, e já se estima que metade dos vírus presentes em habitats marinhos são de ssDNA ou RNA (Delwart & Li, 2012; Labonte & Suttle, 2013; Rosario, Karyna *et al.*, 2012). Apesar de não causarem doenças importantes em humanos, muitos vírus de ssDNA se destacam como patógenos de plantas e de animais domésticos, causando grandes perdas na produção agropecuária (Opriessnig *et al.*, 2020; Rojas *et al.*, 2018; Webb *et al.*, 2020). O grupo mais relevante economicamente constitui o filo *Cressdnaviricota* (domínio *Monodnaviria*). O filo inclui duas classes, oito ordens e onze famílias (*Bacilladnaviridae*, *Circoviridae*, *Geminiviridae*, *Genomoviridae*, *Metaxyviridae*, *Nanoviridae, Naryaviridae, Nenyavirida*, *Redondoviridae*, *Smacoviridae* e *Vilyaviridae*) de vírus que infectam eucariotos e possuem genoma pequeno (2000 a 4000 nucleotídeos) e circular, codificando uma proteína que atua na replicação denominada Rep, que unifica as famílias, características que foram utilizadas para nomear o filo (*circular Rep-encoding ssDNA*) (Krupovic *et al.*, 2020).

Desse modo, considerando as limitações dos algoritmos existentes e as características dos cressdnavírus, como seus genomas pequenos e organizados de forma altamente padronizada, além da confiabilidade na organização taxonômica desse grupo bem estabelecido, esses importantes vírus constituem uma opção interessante para avaliação de metodologias de análises de sequência, principalmente as baseadas em 2D-CNNs na classificação taxonômica. O potencial dasCNNs na classificação taxonômica viral consiste em sua capacidade de identificar automaticamente os mais variados padrões, sugerindo um bom potencial para essa tarefa que não foi ainda bem definida, avaliando também metodologias de aumentação para resolver casos de desbalanceamento de dados.

LITERATURA CITADA

AIEWSAKUN, P.; SIMMONDS, P. The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. **Microbiome,** v. 6, p. 38, 2018.

AUSLANDER, N.; GUSSOW, A.B.; KOONIN, E.V. Incorporating machine learning into established bioinformatics frameworks. **International Journal of Molecular Sciences,** v. 22, p. 2903, 2021.

BAO, Y.; CHETVERNIN, V.; TATUSOVA, T. PAirwise Sequence Comparison (PASC) and its application in the classification of filoviruses. **Viruses,** v. 4, p. 1318-1327, 2012.

BEXFIELD, N.; KELLAM, P. Metagenomics and the molecular identification of novel viruses. **Veterinary Journal,** v. 190, p. 191-198, 2011.

BOLDUC, B.; JANG, H.B.; DOULCIER, G.; YOU, Z.Q.; ROUX, S.; SULLIVAN, M.B. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. **PeerJ,** v. 5, p. e3243, 2017.

DATTA, S.; BUDHAULIYA, R.; DAS, B.; CHATTERJEE, S.; VANLALHMUAKA, V.V. Next-generation sequencing in clinical virology: Discovery of new viruses. **World Journal of Virology,** v. 4, p. 265-276, 2015.

DELWART, E.; LI, L.L. Rapidly expanding genetic diversity and host range of the *Circoviridae* viral family and other Rep encoding small circular ssDNA genomes. **Virus Research,** v. 164, p. 114-121, 2012.

DIETTERICH, T. Overfitting and undercomputing in machine learning. **ACM Computing Surveys,** v. 27, p. 326-327, 1995.

DUTILH, B.E.; VARSANI, A.; TONG, Y.; SIMMONDS, P.; SABANADZOVIC, S.; RUBINO, L.; ROUX, S.; MUÑOZ, A.R.; LOOD, C.; LEFKOWITZ, E.J.; KUHN, J.H.; KRUPOVIC, M.; EDWARDS, R.A.; BRISTER, J.R.; ADRIAENSSENS, E.M.; SULLIVAN, M.B. Perspective on taxonomic classification of uncultivated viruses. **Current Opinion in Virology,** v. 51, p. 207-215, 2021.

EDGAR, R.C.; TAYLOR, J.; LIN, V.; ALTMAN, T.; BARBERA, P.; MELESHKO, D.; LOHR, D.; NOVAKOVSKY, G.; BUCHFINK, B.; AL-SHAYEB, B.; BANFIELD, J.F.; DE LA PEÑA, M.; KOROBEYNIKOV, A.; CHIKHI, R.; BABAIAN, A. Petabase-scale sequence alignment catalyses viral discovery. **Nature,** v. 602, p. 142-147, 2022.

GORBALENYA, A.E.; KRUPOVIC, M.; MUSHEGIAN, A.; KROPINSKI, A.M.; SIDDELL, S.G.; VARSANI, A.; ADAMS, M.J.; DAVISON, A.J.; DUTILH, B.E.; HARRACH, B.; HARRISON, R.L.; JUNGLEN, S.; KING, A.M.Q.; KNOWLES, N.J.; LEFKOWITZ, E.J.; NIBERT, M.L.; RUBINO, L.; SABANADZOVIC, S.; SANFAÇON, H.; SIMMONDS, P.; WALKER, P.J.; ZERBINI, F.M.; KUHN, J.H.; INTERNATIONAL COMMITTEE ON TAXONOMY OF VIRUSES EXECUTIVE, C. The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. **Nature Microbiology,** v. 5, p. 668-674, 2020.

GORBALENYA, A.E.; LAUBER, C. Bioinformatics of virus taxonomy: foundations and tools for developing sequence-based hierarchical classification. **Current Opinion in Virology,** v. 52, p. 48-56, 2022.

GUINDON, S.; DUFAYARD, J.F.; LEFORT, V.; ANISIMOVA, M.; HORDIJK, W.; GASCUEL, O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. **Systematics Biology,** v. 59, p. 307-321, 2010.

HAHN, B.H.; SHAW, G.M.; DE COCK, K.M.; SHARP, P.M. AIDS as a zoonosis: Scientific and public health implications. **Science,** v. 287, p. 607-614, 2000.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. New York: Springer, 2013. ISBN 9781489905185.

IBRAHIM, B.; MCMAHON, D.P.; HUFSKY, F.; BEER, M.; DENG, L.; MERCIER, P.L.; PALMARINI, M.; THIEL, V.; MARZ, M. A new era of virus bioinformatics. **Virus Research,** v. 251, p. 86-90, 2018.

ISLAM, M.A.; KOWAL, M.; JIA, S.; DERPANIS, K.G.; BRUCE, N.D. Position, padding and predictions: A deeper look at position information in cnns. **arXiv**, p. 2101.12322, 2021.

KOONIN, E.V.; DOLJA, V.V.; KRUPOVIC, M.; VARSANI, A.; WOLF, Y.I.; YUTIN, N.; ZERBINI, F.M.; KUHN, J.H. Global organization and proposed megataxonomy of the virus world. **Microbiology and Molecular Biology Reviews,** v. 84, p. e00061-00019, 2020.

KRUPOVIC, M.; VARSANI, A.; KAZLAUSKAS, D.; BREITBART, M.; DELWART, E.; ROSARIO, K.; YUTIN, N.; WOLF, Y.I.; HARRACH, B.; ZERBINI, F.M.J.J.O.V. *Cressdnaviricota*: A virus phylum unifying seven families of rep-encoding viruses with single-stranded, circular DNA genomes. **Journal of Virology,** v. 94, p. e00582-00520, 2020.

LABONTE, J.M.; SUTTLE, C.A. Previously unknown and highly divergent ssDNA viruses populate the oceans. **ISME Journal,** v. 7, p. 2169-2177, 2013.

LAUBER, C.; GORBALENYA, A.E. Partitioning the genetic diversity of a virus family: Approach and evaluation through a case study of picornaviruses. **Journal of Virology,** v. 86, p. 3890-3904, 2012.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature,** v. 521, p. 436-444, 2015.

LI, Z.; LIU, F.; YANG, W.; PENG, S.; ZHOU, J. A survey of convolutional neural networks: Analysis, applications, and prospects. **IEEE Transactions on Neural Networks and Learning Systems,** v. 33, p. 6999-7019, 2022.

MACLOT, F.; CANDRESSE, T.; FILLOUX, D.; MALMSTROM, C.M.; ROUMAGNAC, P.; VAN DER VLUGT, R.; MASSART, S. Illuminating an ecological blackbox: Using High Throughput Sequencing to characterize the plant virome across scales. **Frontiers in Microbiology,** v. 11, p. 2575, 2020.

MARTINEZ, D.A.; NELSON, M.A. The next generation becomes the now generation. **PLOS Genetics,** v. 6, p. e1000906, 2010.

MASKO, D.; HENSMAN, P. **The impact of imbalanced training data for convolutional neural networks**. 2015. Degree Project, First Level (Computer Science). Royal Institute of Technology, Stockholm, Sweden.

MEIER-KOLTHOFF, J.P.; GÖKER, M. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. **Bioinformatics,** v. 33, p. 3396-3404, 2017.

MELO, F. Area under the ROC Curve. In: DUBITZKY, W.*, et al* (Eds.). **Encyclopedia of Systems Biology**. New York: Springer, 2013. p.38-39. ISBN 978-1-4419-9863-7.

MIZUNO, C.M.; RODRIGUEZ-VALERA, F.; KIMES, N.E.; GHAI, R. Expanding the marine virosphere using metagenomics. **PLOS Genetics,** v. 9, p. e1003987, 2013.

MORARU, C.; VARSANI, A.; KROPINSKI, A.M. VIRIDIC - A novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses. **Viruses,** v. 12, p. 1268, 2020.

MUHIRE, B.M.; VARSANI, A.; MARTIN, D.P. SDT: A virus classification tool based on pairwise sequence alignment and identity calculation. **PLOS One** v. 9, p. e108277, 2014.

MURPHY, F.A.; FAUQUET, C.M.; BISHOP, D.H.L.; GHABRIAL, S.A.; JARVIS, A.W.; MARTELLI, A.W.; MAYO, M.A.; SUMMERS, M.D. (Eds.) **Virus Taxonomy. Sixth Report of the International Committee on Taxonomy of Viruses**. Vienna: Springer-Verlag. 1995.

NERI, U.; WOLF, Y.I.; ROUX, S.; CAMARGO, A.P.; LEE, B.; KAZLAUSKAS, D.; CHEN, I.M.; IVANOVA, N.; ZEIGLER ALLEN, L.; PAEZ-ESPINO, D.; BRYANT, D.A.; BHAYA, D.; KRUPOVIC, M.; DOLJA, V.V.; KYRPIDES, N.C.; KOONIN, E.V.; GOPHNA, U. Expansion of the global RNA virome reveals diverse clades of bacteriophages. **Cell,** v. 185, p. 4023-4037, 2022.

OPRIESSNIG, T.; KARUPPANNAN, A.K.; CASTRO, A.; XIAO, C.T. Porcine circoviruses: current status, knowledge gaps and challenges. **Virus Research,** v. 286, p. 198044, 2020.

PAEZ-ESPINO, D.; ELOE-FADROSH, E.A.; PAVLOPOULOS, G.A.; THOMAS, A.D.; HUNTEMANN, M.; MIKHAILOVA, N.; RUBIN, E.; IVANOVA, N.N.; KYRPIDES, N.C. Uncovering Earth's virome. **Nature,** v. 536, p. 425-430, 2016.

PAPPAS, N.; ROUX, S.; HÖLZER, M.; LAMKIEWICZ, K.; MOCK, F.; MARZ, M.; DUTILH, B.E. Virus Bioinformatics. pp. 124-132, *In*: BAMFORD, D.H. & ZUCKERMAN, M. (Eds.). **Encyclopedia of Virology**. 4th Ed., 2021.

PRECHELT, L. Early stopping - But when ? In: MONTAVON, G*., et al* (Ed.). **Neural Networks: Tricks of the Trade: Second Edition**. Berlin: Springer, 2012. p.53-67. ISBN 978-3-642-35289-8.

PRICE, M.N.; DEHAL, P.S.; ARKIN, A.P. FastTree 2–approximately maximum-likelihood trees for large alignments. **PLOS ONE,** v. 5, p. e9490, 2010.

RAJU, R.S.; AL NAHID, A.; CHONDROW DEV, P.; ISLAM, R. VirusTaxo: Taxonomic classification of viruses from the genome sequence using k-mer enrichment. **Genomics,** v. 114, p. 110414, 2022.

REN, J.; SONG, K.; DENG, C.; AHLGREN, N.A.; FUHRMAN, J.A.; LI, Y.; XIE, X.; POPLIN, R.; SUN, F. Identifying viruses from metagenomic data using deep learning. **Quantitative Biology,** v. 8, p. 64-77, 2020.

ROJAS, M.R.; MACEDO, M.A.; MALIANO, M.R.; SOTO-AGUILAR, M.; SOUZA, J.O.; BRIDDON, R.W.; KENYON, L.A.; RIVERA-BUSTAMANTE, R.F.; ZERBINI, F.M.; ADKINS, S.; LEGG, J.P.; KVARNHEDEN, A.; WINTERMANTEL, W.M.; SUDARSHANA, M.R.; PETERSCHMITT, M.; LAPIDOT, M.; MARTIN, D.P.; MORIONES, E.; INOUE-NAGATA, A.K.; GILBERTSON, R.L. World management of geminiviruses. **Annual Review of Phytopathology,** v. 56, p. 637-677, 2018.

ROOSSINCK, M.J.; MARTIN, D.P.; ROUMAGNAC, P. Plant virus metagenomics: advances in virus discovery. **Phytopathology,** v. 105, p. 716-727, 2015.

ROSARIO, K.; BREITBART, M. Exploring the viral world through metagenomics. **Current Opinion in Virology,** v. 1, p. 289-297, 2011.

ROSARIO, K.; DAYARAM, A.; MARINOV, M.; WARE, J.; KRABERGER, S.; STAINTON, D.; BREITBART, M.; VARSANI, A. Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). **Journal of General Virology,** v. 93, p. 2668-2681, 2012.

ROSARIO, K.; DUFFY, S.; BREITBART, M. A field guide to eukaryotic circular single-stranded DNA viruses: Insights gained from metagenomics. **Archives of Virology,** v. 157, p. 1851-1871, 2012.

SHI, M.; LIN, X.D.; TIAN, J.H.; CHEN, L.J.; CHEN, X.; LI, C.X.; QIN, X.C.; LI, J.; CAO, J.P.; EDEN, J.S.; BUCHMANN, J.; WANG, W.; XU, J.; HOLMES, E.C.; ZHANG, Y.Z. Redefining the invertebrate RNA virosphere. **Nature,** v. 540, p. 539-543, 2016.

SIMMONDS, P.; ADAMS, M.J.; BENKŐ, M.; BREITBART, M.; BRISTER, J.R.; CARSTENS, E.B.; DAVISON, A.J.; DELWART, E.; GORBALENYA, A.E.; HARRACH, B.; HULL, R.; KING, A.M.Q.; KOONIN, E.V.; KRUPOVIC, M.; KUHN, J.H.; LEFKOWITZ, E.J.; NIBERT, M.L.; ORTON, R.; ROOSSINCK, M.J.; SABANADZOVIC, S.; SULLIVAN, M.B.; SUTTLE, C.A.; TESH, R.B.; VAN DER VLUGT, R.A.; VARSANI, A.; ZERBINI, F.M. Consensus statement: virus taxonomy in the age of metagenomics. **Nature Reviews Microbiology,** v. 15, p. 161, 2017.

SIMMONDS, P.; ADRIAENSSENS, E.M.; ZERBINI, F.M.; ABRESCIA, N.G.A.; AIEWSAKUN, P.; ALFENAS-ZERBINI, P.; BAO, Y.; BARYLSKI, J.; DROSTEN, C.; DUFFY, S.; DUPREX, W.P.; DUTILH, B.E.; ELENA, S.F.; GARCÍA, M.L.; JUNGLEN, S.; KATZOURAKIS, A.; KOONIN, E.V.; KRUPOVIC, M.; KUHN, J.H.; LAMBERT, A.J.; LEFKOWITZ, E.J.; ŁOBOCKA, M.; LOOD, C.; MAHONY, J.; MEIER-KOLTHOFF, J.P.; MUSHEGIAN, A.R.; OKSANEN, H.M.; PORANEN, M.M.; REYES-MUÑOZ, A.; ROBERTSON, D.L.; ROUX, S.; RUBINO, L.; SABANADZOVIC, S.; SIDDELL, S.; SKERN, T.; SMITH, D.B.; SULLIVAN, M.B.; SUZUKI, N.; TURNER, D.; VAN DOORSLAER, K.; VANDAMME, A.-M.; VARSANI, A.; VASILAKIS, N. Four principles to establish a universal virus taxonomy. **PLOS Biology,** v. 21, p. e3001922, 2023.

SOFAER, H.R.; HOETING, J.A.; JARNEVICH, C.S. The area under the precision-recall curve as a performance metric for rare binary events. **Methods in Ecology and Evolution,** v. 10, p. 565-577, 2019.

SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; SUTSKEVER, I.; SALAKHUTDINOV, R. Dropout: A simple way to prevent neural networks from overfitting. **Journal of Machine Learning Research,** v. 15, p. 1929-1958, 2014.

SUTTLE, C.A. Viruses in the sea. **Nature,** v. 437, p. 356, 2005.

TAJUDEEN, Y.A.; OLADIPO, H.J.; YUSUF, R.O.; OLADUNJOYE, I.O.; ADEBAYO, A.O.; AHMED, A.F.; EL-SHERBINI, M.S. The need to prioritize prevention of viral spillover in the Anthropopandemicene: A message to global health researchers and policymakers. **Challenges,** v. 13, p. 35, 2022.

TISZA, M.J.; PASTRANA, D.V.; WELCH, N.L.; STEWART, B.; PERETTI, A.; STARRETT, G.J.; PANG, Y.S.; KRISHNAMURTHY, S.R.; PESAVENTO, P.A.; MCDERMOTT, D.H.; MURPHY, P.M.; WHITED, J.L.; MILLER, B.; BRENCHLEY, J.; ROSSHART, S.P.; REHERMANN, B.; DOORBAR, J.; TA'ALA, B.A.; PLETNIKOVA, O.; TRONCOSO, J.C.; RESNICK, S.M.; BOLDUC, B.; SULLIVAN, M.B.; VARSANI, A.; SEGALL, A.M.; BUCK, C.B. Discovery of several thousand highly diverse circular DNA viruses. **eLife,** v. 9, p. e51971, 2020.

WARDHANI, N.W.S.; ROCHAYANI, M.Y.; IRIANY, A.; SULISTYONO, A.D.; LESTANTYO, P. **Cross-validation metrics for evaluating classification performance on imbalanced data**. 2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA). Tangerang, Indonesia. 23-24 Oct. 2019, 2019. 14-18 p.

WEBB, B.; RAKIBUZZAMAN, A.; RAMAMOORTHY, S. Torque teno viruses in health and disease. **Virus Research,** v. 285, p. 198013, 2020.

CHAPTER 1

TAXONOMIC CLASSIFICATION OF CRESS-DNA VIRUSES USING 2D CONVOLUTIONAL NEURAL NETWORKS

Gomes RAL, Zerbini FM. Taxonomic classification of CRESS-DNA viruses using 2D convolutional neural networks. *Journal of Virological Methods*, submitted.

Taxonomic classification of CRESS-DNA viruses using 2D convolutional neural networks

Ruither A. L. Gomes[1,2], F. Murilo Zerbini[1,2*]

[1]Dep. de Fitopatologia, [2]National Institute for Science and Technology on Plant-Pest Interactions, Universidade Federal de Viçosa, Viçosa, MG, 36570-900, Brazil;

*Corresponding author:

Phone: (+55-31) 3612-2423; E-mail: zerbini@ufv.br

ABSTRACT

Taxonomy, defined as the classification of different objects/organisms into defined stable hierarchical categories (taxa), is fundamental for proper scientific communication. In virology, taxonomic assignments based on sequence alone are now possible and their use may contribute to a more precise and comprehensive framework. The current major challenge is to develop tools for the automated classification of the millions of putative new viruses discovered in metagenomic studies. Among the many tools that have been proposed, those applying machine learning (ML), mainly in the deep learning branch, stand out with highly accurate results. One ML tool recently released that uses k-mers, VirusTaxo, was the first one to be applied with success, 93% average accuracy, to all types of viruses. Nevertheless, there is a demand for new tools that are less computationally intensive. Viruses classified in the phylum *Cressdnaviricota*, with their small and compact genomes, are good subjects for testing these new tools. Here, we tested the usage of 2D convolutional neural networks for the taxonomic classification of cressdnaviricots, also assessing the effect of data imbalance and two augmentation techniques by benchmarking against VirusTaxo. We were able to get perfect classification during k-fold test evaluations for balanced taxas, and more than 98% accuracy in the final pipeline tested for imbalanced datasets. The mixture of augmentation on more imbalanced groups and no augmentation for more balanced ones achieved the best score in the final test. These results indicate that these architectures can classify DNA sequences with high precision.

INTRODUCTION

Taxonomy, defined as the categorization of objects or beings in an organized, hierarchical way, is a core part of science. Taxonomy allows scientists to better communicate with each other by linking objects or beings to stably named categories (taxa) (Gorbalenya, Lauber, and Siddell, 2019). Viruses are probably the most diverse and ubiquitous biological entities on the planet (Paez-Espino et al., 2016). Viral discovery during most of virology's history was based on purifying viruses from infected hosts or cell cultures and on PCR-based sequencing of viral genomes, with taxonomic assignments being made individually by different research groups for each new virus based on phenotypic characteristics of the infection and molecular properties (Datta et al., 2015). As stated by Fauquet (1999), "As in other biological systems, virus classification is an approximate and imperfect exercise. Like any other type of classification, it is a totally artificial and human-driven activity". In recent years, major changes in the taxonomic framework were made towards a more direct genome-based classification (Simmonds et al., 2023).

As a consequence of the exponential increase in viral sequence data from high-throughput sequencing (HTS) of metagenomic samples, the International Committee on Taxonomy of Viruses (ICTV) decided, in 2017, to start accepting metagenomic generated sequences (after checking for completeness and quality) as true representatives of viruses (Simmonds et al., 2017). Since then, several taxa (species, genera, families and one order) have been created based exclusively on metagenomic-derived sequences (Koonin and Yutin, 2020; Krupovic et al., 2016). Additionally, the ICTV invited the virology community to assist in the development of new automated classification methods to deal with large datasets (Simmonds et al., 2017).

Different computational classification tools have been proposed in the two last decades to help dealing with the new reality in viral biodiversity. Most of these tools are based on sequence clustering methodologies using demarcation thresholds that separate taxonomic levels using sequence divergence between one or more conserved genes in a group. These methods are usually applicable for classification at the family and genus ranks (Pappas et al., 2021).

Machine learning (ML) algorithms have already lead to incredible developments for different types of data analyses. In bioinformatics, they have been successfully applied to a broad range of fields, from the prediction of RNA or protein structure to phylogenetic and biological network analyses. In taxonomy, ML methods are mostly based on extraction and selection of different features from datasets of genomic sequences, and development of prediction models that can identify the taxonomic rank of new sequences without the need for direct comparison with all members of a selected group using sequence alignments, as done by most other tools. Thus, ML algorithms are much less computationally intensive than sequence alignment-based methods, allowing the trained ML models to be used in huge query datasets than would be unfeasible otherwise (Auslander, Gussow, and Koonin, 2021).

When dealing with text, audio, image, and video recognition, convolutional neural networks (CNNs), a branch of deep learning (DL), are powerful tools due to their ability to extract patterns from large input datasets (Li et al., 2022). In recent studies proposing new bioinformatic tools, CNNs have been applied with highly accurate results in some fields, such as the identification of enhancer-promoter interactions (Min et al., 2021), the prediction of DNA-binding proteins (Barukab et al., 2022; Zhang et al., 2019), and the detection of patterns in DNA methylation (Zeng and Gifford, 2017). Specifically for viral classification, some examples are Viral Genome Deep Classifier, DeepVirFinder, EdeepVPP and EdeepVPP-hybrid (Dasari and Bhukya, 2022; Fabijańska and Grabowski, 2019; Ren et al., 2020), and BERTax, a method that uses natural language processing and achieved 95% accuracy at the phylum level

of all four "superkingdoms" defined by the authors (archaea, bacteria, eukaryota, and viruses) (Mock et al., 2022). With the exception of BERTax, the algorithms achieved highly accurate training results only in specific datasets. Two of them (DeepVirFinder and EdeepVPP) were only able to predict if the input was a viral sequence, not inferring actual taxonomic classification, and the third one (Viral Genome Deep Classifier) was only trained to classify subtypes of five human viruses that had more than 10 similar sequences available. And even in the case of BERTax, in spite of its good accuracy at the phylum level, classification at the genus level achieved only 75% accuracy, which is inferior to other tools. Another common aspect of these tools is the use of text techniques, such as BioVec (Asgari and Mofrad, 2015) or BERT (Devlin et al., 2018), as done in BERTax, as the representation of the genomic sequence. Moreover, most of the times these were applied to one-dimensional (1D) CNNs, with only a few works using two-dimensional (2D) CNNs, and fewer applying to the taxonomic classification.

For classification into taxonomic ranks, ML algorithms have been developed that use assembled contigs as inputs to classify the sequence with high accuracy. The most recently developed, VirusTaxo (Raju et al., 2022), is based on a simple but memory-intensive detection of unique sequence fragments of length *k* (k-mers). VirusTaxo was the first tool trained with almost all virus genera available (of both RNA and DNA viruses), removing only those that had unique sequences (named singletons). The algorithm obtained an average accuracy of 93% during training. VirusTaxo was benchmarked against two commonly used tools based on k-mers, Kraken2 (Wood, Lu, and Langmead, 2019) and CLARK (Ounit et al., 2015), and also against DeepVirFinder, a CNN-based viral sequence predictor. It got similar results as DeepVirFinder for the identification of viral contigs in metagenome datasets, while Kraken2 and CLARK detected almost 100 times fewer viral sequences in the tested samples. Also, VirusTaxo achieved similar results as the other two k-mer-based tools while predicting

taxonomic classification at the genus level from a dataset of partial and fully assembled contigs from the virus SARS-CoV-2.

The existing tools have different limitations for the identification and taxonomic classification of viral sequences. To date, only VirusTaxo can achieve the identification of viral sequences in metagenomic datasets and also predict the taxonomic classification of those sequences with confidence. However, this tool also has limitations, such as high RAM usage (for common computers, >24 GB) and the fact that k-mer-based approaches may be too sensitive to data-specific characteristics such as mutation, recombination or region-specific rates of evolution (Gorbalenya and Lauber, 2022). Moreover, VirusTaxo was only tested to predict novel (unseen) genomes by removing one species from the training dataset and predicting it at the correct genus. However, the sequence similarity is usually very high among members of the same genus (more than 96% in the tested case).

Different approaches using CNN for DNA sequences in various fields achieved high accuracy, even in the identification of viruses, as seen in DeepVirFinder, but there is a lack of viral taxonomic classification tools using 2D-CNNs. This reinforces the need for the development of additional tools for viral taxonomic classification.

One interesting group of viruses that were shown to have greater diversity than previously thought are the DNA viruses with small, circular, single-stranded genomes. These viruses do not seem to cause disease in humans but are known to cause major losses in crops and livestock, and in consequence, have been studied in detail through the years (Rosario, Duffy, and Breitbart, 2012). Among these viruses, some families comprise a distinct group that encode a conserved replication-initiation protein (named Rep) involved in the initiation step of rolling-circle replication, an important replication mechanism for circular DNA genomes also used in some plasmids. These families were informally named CRESS-DNA (circular, Rep-encoding ssDNA) viruses, and were recently classified in a new phylum designated

Cressdnaviricota (Krupovic et al., 2020). The family *Geminiviridae*, which includes a large number of economically important plant viruses that cause severe diseases in tropical and subtropical reagions worldwide (Rojas et al., 2018), is classifed in this phylum.

The objective of this work was to evaluate the capacity of 2D-CNNs to be used in the taxonomic classification of viruses, using members of the phylum *Cressdnaviricota* as test subjects, and test its capacity against the best tool currently available.

MATERIALS AND METHODS

Computational resources

All codes and analyses in this work were done using Python 3 (python.org/doc/). The models and functions used are from Keras v. 2.9.0 (github.com/keras-team/keras), TensorFlow v. 2.9.1 (TF, (Abadi et al., 2016) and SKlearn v.1.2.2 (Pedregosa et al., 2011).

Data collection

The virus sequences used in the training step of the model were obtained from the ICTV's Virus Metadata Resource (VMR, ictv.global/vmr), a taxonomic list of well-characterized and defined virus isolates from all classified species, using the 08/31/2022 release (VMR_20-190822_MSL37.2.xlsx). From this file, GenBank accession numbers of all *Cressdnaviricota* members were extracted and used to download the genomic sequences from the database using NCBI's Entrez, available in the Biopython library (Cock et al., 2009). Also, all *Cressdnaviricota*-associated information available was downloaded from the NCBI Genome Data Hub (ncbi.nlm.nih.gov/data-hub/genome/) using Entrez. Those accessions that were already used in the model training set, or that were Reference Sequences from NCBI (started with "NC_") and had 100% identity and coverage with model sequences, were removed. The

remaining sequences were used as the final test dataset forbenchmarking the best model using aprediction algorithm to compare with VirusTaxo.

Convolutional Neural Networks

We tested supervised (labeled data) 2D-CNNs, which are commonly composed of different types of layers which may change the data in specific ways. We evaluated different architectures with common layers available to usage in 2D-CNNs with TF: convolutional, pooling, dropout and dense layers, in a "manual search", trying each architecture with hyperparameters similar to those found on CNN sequence classifiers that yielded good results, and then refining them by testing variations and combinations. In our tests, the rectified linear unit (ReLU) function (Nair and Hinton, 2010) was used as the activation function for all layers (convolutional, pooling and dense), with exception of the last dense layer that used softmax (Araújo et al., 2017). Padding was selected to "same" and stride was selected to default. Training occurs by updating the network trainable parameters by applying a stochastic gradient and back-propagation to reduce loss (LeCun, Bengio, and Hinton, 2015), which in our case was done by the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-4} and other parameters as default, , and using the categorical cross-entropy loss function, which "evaluates the difference between the probability distribution obtained from the current training and the actual distribution" (Li et al., 2022), and that is often applied in the case of multiclass classification.

Data encoding and input standardization

One-hot encoding is a simple and practical way of encoding textual data in a numerical format, used in different computation approaches. In DNA and protein codification, each possible base/amino acid is converted to a fixed size vector, with zeros in all positions except

in one (hence the method's name), corresponding to that letter (base or aminoacid letter symbol) in that position. In our case, for DNA we codified the nucleotides A, T, C and G as [1,0,0,0], [0,1,0,0], [0,0,1,0] and [0,0,0,1], respectively, while for every other possible DNA representation we used the vector [0,0,0,0]. So, for example, when encoding a DNA sequence of length L, the algorithm transforms the sequence in a Lx4 matrix, where each line represents the position of each DNA base in the sequence and each column represents which letter is using "one" (Figure 1).

A	1	0	0	0
G	0	0	0	1
G	0	0	0	1
C	0	0	1	0
-	0	0	0	0
T	0	1	0	0
A	1	0	0	0
T	0	1	0	0
N	0	0	0	0
T	0	1	0	0

Figure 1. Example of the one-hot encoding for the sequence AGGC-TATNT, where A = [1,0,0,0], T = [0,1,0,0], C = [0,0,1,0], G = [0,0,0,1] and any other character = [0,0,0,0].

In some works, this matrix is applied to a 1D-CNN, where each nucleotide position (column) goes through individual channels which search for patterns within each type of nucleotide. In our approach, we use the less common option of using this matrix as an image

and passing it through a single-channel 2D-CNN. Thus, the matrix is used as one whole thing, with each filter seeking patterns in all four nucleotides at the same time.

Before the coding part of the sequences, an important step is formatting the input sequences. Initially, as some viruses have multipartite genomes (where their genomic information exists in multiple segments encapsulated in different viral particles), we concatenated the segments into a single molecule, as CNNs kernels are not positionally limited in consequence of moving entirely through the matrices looking wherever the patterns may be. Moreover, CNNs require a fixed-size matrix as input, which has been done differently by each approach, such as completing with zeros/gaps the remaining size or truncating the sequences in specific sizes. In our case, seeking to use complete genomes in all inputs, this was done based on the idea seen in copy-paste augmentation methodology for image recognition, where copies of one object are added in the same image, and which has been shown to be a very effective and robust augmentation method in other tests (Ghiasi et al., 2021; Pappagari et al., 2021). Thus, every sequence had its size completed to the fixed input length by adding copies of itself to its end portion until it reached the desired length, which in this work was set to 15,000 positions/nucleotides (15,000 x 4 matrix as input). In the case of the longest *Cressdnaviricota* genomes, those of viruses from the genus *Nanovirus* (8 segments with aproximately 1,000 nucleotides each), this lead to almost two copies of the concatenated genome, allowing a considerable margin for variation.

Data imbalance

The composition of the dataset is arguably the most important part when developing a prediction model, since the model's ability to predict is defined and limited by the dataset. The dataset is considered to be imbalanced when the different categories have data distributions that are considerably different amongst themselves. For example, if almost all samples in a dataset

with two groups belong to one of them, the model could get biased in classifying every sample as the largest group, as it would get good accuracy by that. Different methods have been proposed to solve this problem, mostly based on different ways of undersampling the largest dataset or oversampling the smallest dataset (Lin et al., 2017; Masko and Hensman, 2015).

In our study case there is an imbalance in different taxonomic ranks, mainly caused by the bias on researchers studing more economically important pathogens. The main considerable example is the genus *Begomovirus*, belonging to the family *Geminiviridae*, which includes several hundred species, with more than 600 genomes in the VMR file. Seeking to reduce this number to values closer to other genera, we applied CD-Hit (Fu et al., 2012) specifically for the training/curated dataset to cluster all begomovirus sequences with more than 70% sequence identity, with other parameters set to default. After this clustering, we only used one sequence as representative of each cluster, reducing the number of begomovirus sequences from more than 600 to 118.

We also checked the effect of two techniques to solve oversampling during the tests, trying to balance small datasets (such as at the genus level, where many genera only have a small number of sequences associated with them) by creating variations in copies of the existing data. In the first tecnique, mutations of a random percentage of bases are applied to a specific number of copies of existing sequences, similar to what is suggested by Busia *et. al.* (2018), randomly mutating 1-2% of different sequences until the groups are equivalent in data. The second approach was the addition of a sequence fragment, randomly selected from a pool, to each border of sequence copies, e.g. inittialy splitting the sequences of each group into fragments that are held in the same pool and are randomly sampled. This is also similar to the idea of copy-paste augmentation, but with fragments of the viral sequences being added to the borders to increase the data variability. Considering that most viruses have compact genomes

and most portions of the genome normally carry information associated with that sequence, it is unlikely that the fragments will not carry new information/patterns to the sequence.

After defining the main methodologies, functions and metrics, important hyperparameters (number and types of layers, number of kernels, kernel size, dropout, max-pooling windows size, number of neurons in dense layers) were evaluated by testing some previously used architectures for DNA sequence classification with CNNs. In one recent example (Dasari and Bhukya, 2022), a CNN model for the detection of human viral pathogens on human metagenomic datasets applied a 1D-CNN architecture composed of three groups of layers for extracting more complex features in every layer. Each group of layers was applied sequentially: a 1D convolutional layer (32 kernels, kernel size 7; 8 kernels, size 4; 8 kernels, size 3), a dropout layer (0.2 in all three groups), and a max-pooling layer (kernel size 2, stride 2 in all three groups). Then, the flattening layer condensed the information into a vector used as input to a dense layer with 32 neurons. Finally, an output layer with two neurons released the classification (viral or non-viral sequence) with 99% accuracy on training.

A different approach was applied with good results by Min et al. (2021) using larger window sizes. In this case, the 1D convolutional layer and the max-pooling layer used filter sizes ranging from 40 to 60 and from 20 to 30, respectively. These larger kernels are interesting because they appear to be similar to the optimal size of k-mers used in VirusTaxo, which is 21 nucleotides. Also when converted into nucleotide sequences, the sizes of the average canonical protein backbone fragments used in databases (4 to 14 amino acids) (Baeten et al., 2008) would correspond to a range of 12 to 42 nucleotides. Based on these different architectures, we tried applying them directly, combined, and with different variations as seen in other studies.

RESULTS AND DISCUSSION

Model architecture

For testing the different possible model architectures, we initially used as datasets the first branching of *Cressdnaviricota*, that between the classes *Alfiviricetes* and *Repensiviricetes*. This allowed us to test without using augmentation techniques, as the proportion of members between the groups is near 1:2 (252 and 474 members for *Alfiviricetes* and *Repensivirices*, respectively), which other CNN studies have shown not to be detrimental to the results (Qu et al., 2020). As this dataset comprises all sequences that would be used for training and are defined in only two groups, it seemed practical for evaluating the performance effect of different architectures on it.

The performance was evaluated during training using AUC-PR applying the stratified *K*-fold approach ($K = 5$), where for each fold the AUC-PR and the F1 score were calculated based on the validation set and used to get the final average. Also, the early stopping function was set to monitor the validation loss, and the patience was set initially to 5. However, after testing higher values, we found that even after some dozens of epochs the loss could still drop and classify more sequences correctly. So from almost the beginning we set the patience to 100, with a higher number of maximum epochs to accommodate large trainings, and started getting better results after a few hundred epochs. Finally, we initially used a batch size of 100, but in the first trials, we found that using a lower value tended to improve the accuracy so it was eventually set to 2, which gave better results and agreed with the results of Masters & Luschi (2018), at the cost of longer training time.

The first architecture tested was the one used by Dasari and Bhukya (2022). We applied it directly, adapting only the filter sizes to 2D convolution (7x7, 4x4 and 3x3 convolutional kernel sizes, and 2x2 max-pooling sizes). This configuration achieved an average AUC-PR of

93.131 with standard deviation (STD) of 1.498, and average F1 score of 86.662 and STD of 3.302, which meant fewer than a hundred sequences being wrongly classified. Nevertheless, the validation AUC and especially the loss graphs are highly scattered (Figure 2), indicating that this architecture cannot find distinguishing patterns in the data.

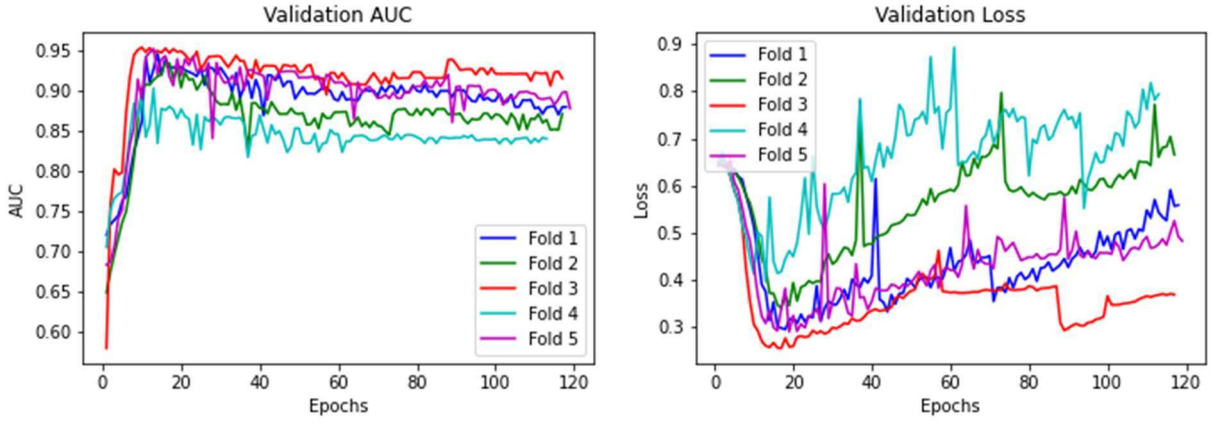

Figure 2. Training metrics (validation AUC-PR and loss) of a 2D-CNN for sequence-based viral classification in the phylum *Cressdnaviricota*, using a similar architecture than that of Dasari and Bhukya (2022).

Besides other changes that did not affect the results, we tried using the larger filter sizes for convolutional and max-pooling layers, as done by Min et al. (2021). Thus we changed the sizer of all three layers using 40x4 filters (only four columns as the sequence vector has only four positions) in the first two convolutional filters, and 30x4 in the last one, and changing the max-pooling filter to 20x4, seeking to find higher patterns in convolutions and reduce the dimensions, keeping only what is important (higher values). With these changes the model achieved an average AUC-PR of 99.4 (STD of 0.547) and average F1 score of 96.774 (STD of 2.585) (Figure 3), with only two dozen sequences being wrongly classified, thus showing that

larger filters increased the prediction power of the CNN. However, even with good AUC and F1 scores the model made wrong predictions, indicating that there was still space for further improvement.

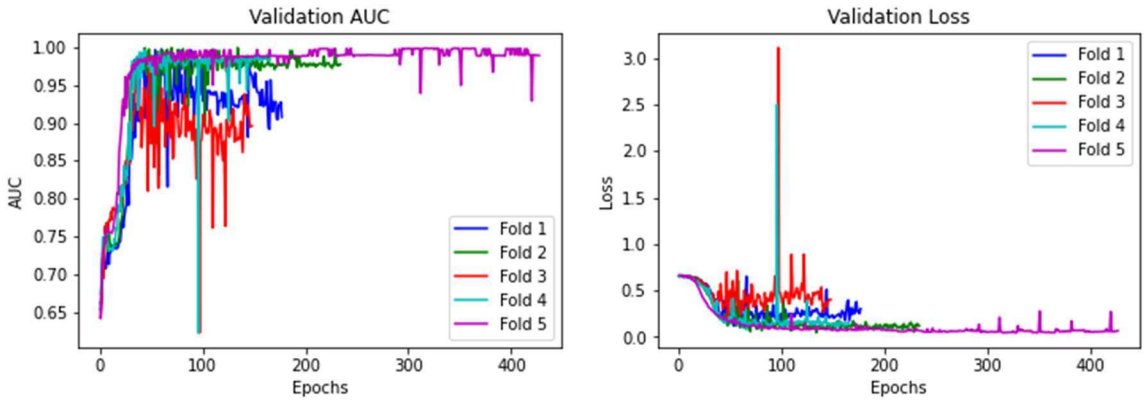

Figure 3. Training metrics (validation AUC-PR and loss) of a 2D-CNN for sequence-based viral classification in the phylum *Cressdnaviricota*, using the architecture proposed by Dasari & Bhukya (2022) with larger filters (convolutional kernels, 40x4 and 30x4; max-pooling, 20x4), as proposed by Min et al. (2021).

We then tried different combinations of hyperparameters, changing the number of layers, the number of convolutional kernels and their sizes, layer orders, dropout percentages, and max-pooling filter sizes, changing them one at a time, maintaining those configurations that enhanced the model's performance and trying additional variations from them. After extensive testing with different architectures, excellent performance was achieved by reducing the max-pooling column size from 4 to 1 (where the dimensions are merged for each nucleotide, maintaining the matrix width). This configuration achieved an average F1-score of 98.893 (STD of 0.709) and AUC-PR of 99.762 (STD of 0.364). The validation graphs (Figure 4) show that the curves quickly attained good values and started flattening, with more punctual peaks compared to the previous configurations, indicating more stability during training.

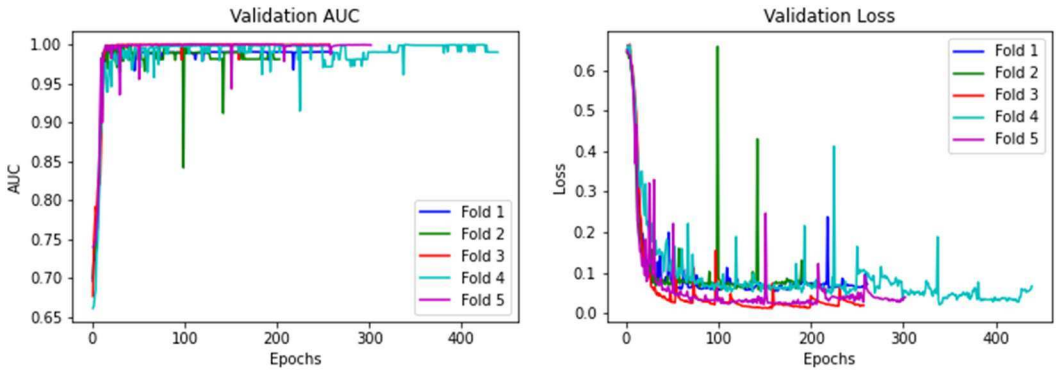

Figure 4. Training metrics (validation AUC-PR and loss) of a 2D-CNN for sequence-based viral classification in the phylum *Cressdnaviricota*, using an architecture incorporating elements from Dasari and Bhukya (2022) and Min et al. (2021) and the width of max-pooling layers reduced from 4 to 1.

Another change that enhanced the model's performance was removing one group of convolutional layers and changing the convolutional kernel size. The first CNN that stood out used 16 150x4 filters in the first layer, followed by a max-pooling layer with a 50x1 filter and a dropout layer of 0.5. The next convolutional layer had 32 50x4 filters, a max-pooling of 25x1 and a dropout also of 0.5. This was followed by the flatten layer, a dense layer with 64 neurons, another dropout layer with 0.4, and finally the output layer releasing the prediction vector. With this architecture, the model was able to achieve an average AUC-PR of 99.812 (STD of 0.369) and F1 score of 99.862 (SDT of 0.275), with only one sequence wrongly predicted. The validation graphs (Figure 5) display patterns of flattening and punctual peaks, however more uniform between the five folds, and also with values of AUC varying mostly near 0.999. The loss graph was more constant between folds, with considerable overlapping.

A final change in the model (64 instead of 16 convolutional kernels in the first convolutional layer; Figure 6) achieved perfect (1.0) AUC and F1 scores, with 0 STD and no wrongly classified sequences. A slightly worse result (two wrongly classified sequences) was obtained in a sequential test. Seeking to better elucidate the differences between the two architectures, we used repeated stratified k-fold cross-validation with 5 repetitions using 5 folds

each, and all the 25 accuracies and F1 scores were used to apply a paired t-tes. We obtained *p*-values of 0.4587 and 0.0052, respectively, showing that there are statistically significant differences between the models for the F1 score, with the 64-kernel architecture yielding better results.

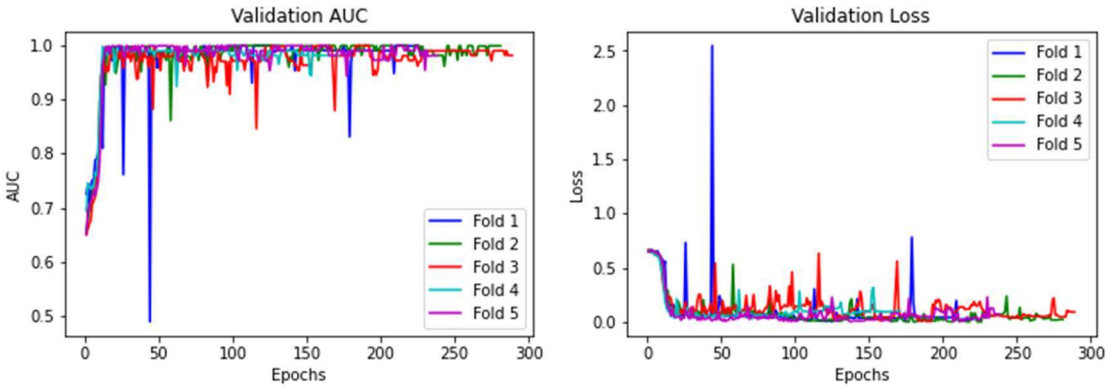

Figure 5. Training metrics (validation AUC-PR and loss) of a 2D-CNN for sequence-based viral classification in the phylum *Cressdnaviricota*, using an architecture incorporating elements from Dasari and Bhukya (2022) and Min et al. (2021), but with the width of max-pooling layers reduced from 4 to 1, one less group of layers, and bigger filter sizes.

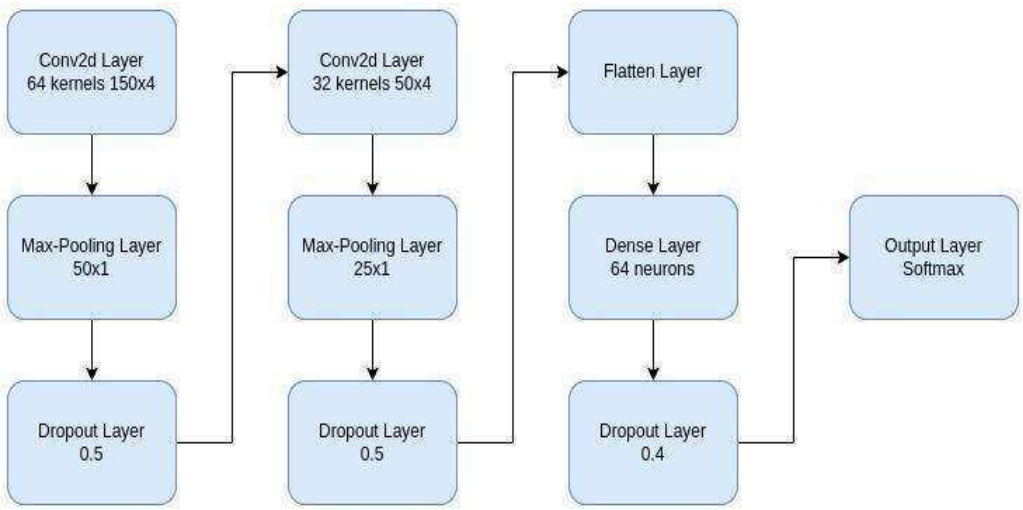

Figure 6. The architecture of the CNN with best results during training, where in one test it classified all sequences correctly.

With these metrics for the two latter architectures, we chose to use the one with 64 initial convolutional kernels to test for lower taxonomic ranks. First we trained the model for the family *Circoviridae*, which includes two genera, *Circovirus* and *Cyclovirus*, with 49 and 52 sequences, respectively. The model achieved high values, with an average AUC-PR of 99.902 (STD of 0.120), and F1 score of 97.995 (STD of 2.456), indicating that it worked well in balanced data even for the lower ranks.

However, imbalanced data composes most of the families in the phylum, so we also trained the model in one imbalanced family of metagenomic characterized viruses, the *Genomoviridae*. In this case the model achieved an AUC-PR of only 84.209, while the F1 score was 89.369. The model made mostly correct predictions but with low confidence in the probabilities, with 24 wrongly classified sequences in a dataset of almost 100 sequences. The low AUC-PR and F1 scores can be expected considering that one genus has only three sequences, seven genera have about 10 sequences in average, one genus has 50+ sequences, and one genus (*Genomovirus*) has 190+ sequences. This imbalance, which is also observed in most of the other families and is a common sampling problem in virus taxonomy, is likely the main factor responsible for the lower values obtained for the two metrics.

One important characteristic of ML algorithms is the division of the dataset into at least two groups, one for training and one for testing. Thus, when only one sequence is available in a genus (known as singletons), this genus was removed from the training set. This was the case for 26 genera. To avoid ignoring these genera, each input was also analyzed by BLAST (Altschul et al., 1990) against these singletons.

Pipeline development

We decided to use a hierarchical pipeline for classification which is based on a series of models parsing each input by classifying them from the uppermost rank of phylum down to

genus, which means predicting four taxonomic ranks (or three in some cases, see below). This hierarchical methodology seeks to reduce the sampling problem shown by the small number of species in some genera, trying to confine the bias to the related family model with imbalance.

The pipeline first receives an input matrix from each sequence by one-hot encoding (as seen in Figure 7 and described in the section *Data encoding and input standardization* in the Material and Methods). This matrix is then used for prediction at all ranks, where the predicted rank is used to select the model to be used for the next rank prediction. For example, the first model separates the *Cressdnaviricota* members between those classified in the classes *Arfiviricetes* or *Repenviricetes,* and marks the prediction type as done by "model". Once the class has been predicted, the matrix is then used to obtain the order predictions by the second model (in the case of only one existing group there is no model, thus the prediction type is marked as "single" and goes directly to the existing rank), repeating this process until reaching the genus classification. After the genus prediction, the pipeline runs a BLAST analysis against all sequences from the other genera in the predicted family, checking if only sequences from the predicted genus have alignment similarity. When there are BLAST hits for the other genera (including the singletons), it uses the hit with the highest identity and coverage as the output prediction.

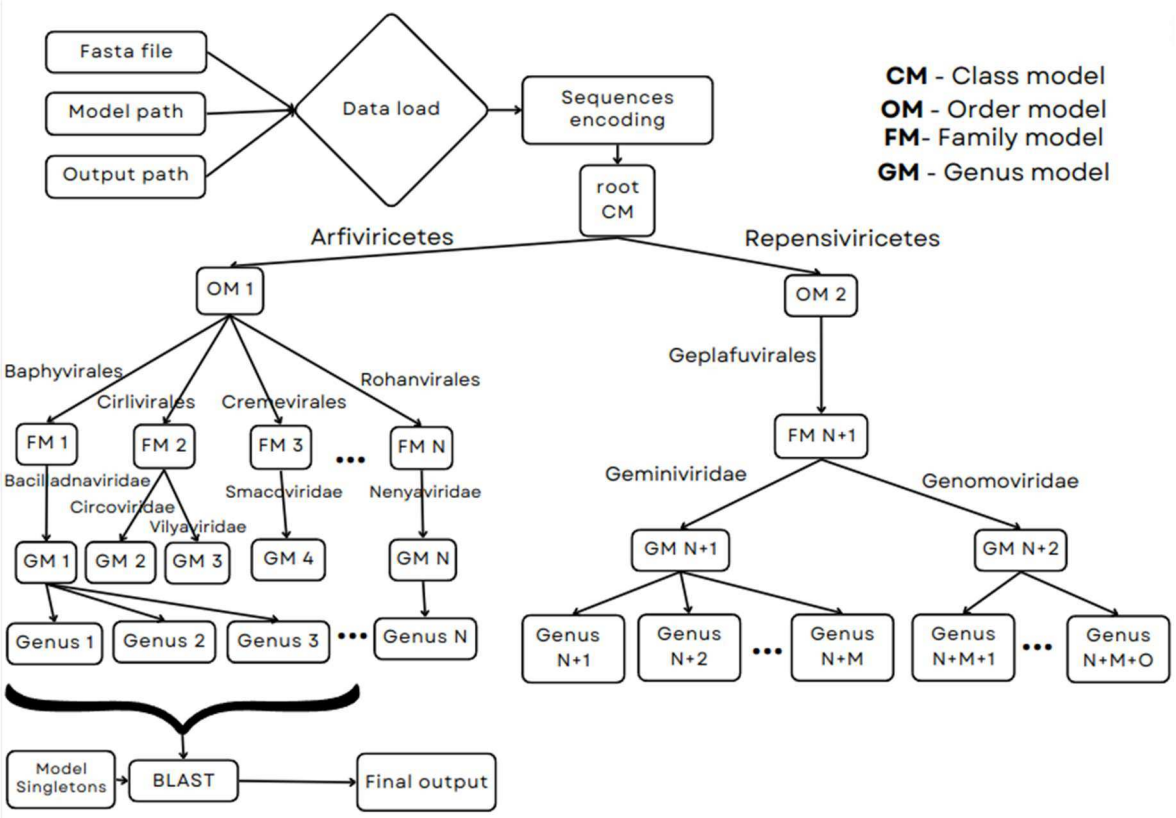

Figure 7. Classification pipeline of the prediction script based on hierarchical taxonomic classification, where every level in the viral taxonomy is predicted separately and the result is used in the next prediction. The user provides the fasta file with the sequences to be tested and uses the main model files (with the best results until now) to get the predictions, which are saved in .csv files in the output directory. After predictions for each taxonomic level using the respective model, the algorithm uses all genera in the predicted family in a BLAST test to confirm the predictions using the sequence similarity and to correct wrong predictions at the genus level, as genera tend to show more imbalance in the datasets.

Pipeline benchmarking

The taxonomic classification capability of our model was first evaluated by comparing it with that of VirusTaxo to classify *Cressdnaviricota* sequences available on the NCBI's genome datahub. After removing the used accession numbers and reference sequences that had 100% identity and coverage using BLAST, we selected only the sequences that were classified at the genus level, which left us with 4,170 sequences for testing. For comparison with VirusTaxo, we used its database creation function to build a custom dataset with all sequences

used in our model training, using a k-mer size of 21, which is the value indicated by the program for DNA sequences.

The models tested here were built using the two best architectures found above with two augmentation techniques (percentage augmentation, PercAug, and border augmentation, BordAug), with no augmentation (NoAug), and also with a combination of the phylum model using NoAug while the other models used BordAug or PercAug, totaling 10 different pipeline ensembles tested against VirusTaxo. Also, after the predictions, all results were parsed and sequences that had 100% identity and coverage were again removed, which led to 4,051 sequences remaining that were never "seen" by the models. The vast majority (3,647) of these sequences belong to the genus *Circovirus*, while some other genera had only one or two sequences, such as *Curtovirus*, *Grablovirus*, *Capulavirus*, *Topilevirus*, *Bacilladnavirus*, *Gemykolovirus* and *Gemykrogvirus*. As a consequence of this small number of sequences, these genera and some others with few members (or that were misclassified by all tests) were not placed in the final result calculations below.

All tested pipelines got metrics that were similar to those obtained with VirusTaxo for the most populated genera, but most of them misclassified fewer sequences than VirusTaxo (Table 1, Table 2). The pipeline composed of a NoAug model for phylum and BordAug for the remaining models achieved the highest accuracy (98.56%) and only misclassified 58 viruses, while VirusTaxo misclassified 72 viruses.

Table 1. Classification results for the benchmarking dataset tested with 10 different model ensembles developed here and with VirusTaxo, where positives are sequences classified correctly, while negatives represent the misclassified sequences.

Method/ Classification	Total positives	Total negatives	Average accuracy* (%)
VirusTaxo	3,979	72	98.22
NoAug	3,989	62	98.46
PercAug	3,969	82	97.97
BordAug	3,987	64	98.42
NoAug + PercAug	3,969	82	97.97
NoAug + BordAug	3,987	64	98.42
NoAug (64)	3,982	69	98.29
PercAug (64)	3,984	67	98.34
BordAug (64)	3,987	64	98.42
NoAug + PercAug (64)	3,989	62	98.46
NoAug + BordAug (64)	3,993	58	98.56

*Not taking into consideration three genera that were misclassified in all tests

The results also demonstrate how augmentation by itself had a small and not guaranteed beneficial effect on the results, with a worse result in one case, and how by mixing methodologies it was possible to achieve better metrics. Another noticeable result is that some sequences previously classified as circoviruses in the NCBI database got higher identity and coverage with sequences in the genus *Cyclovirus*. These sequences were classified as mistakes, but with the possibility that the error actually took place in the original classification, as they did not get BLAST hits with the circoviruses used in the model. Furthermore, all sequences in

the genera *Drosmacovirus* and *Glamdringvirus*, and eight sequences in the genus *Gemycircularvirus* were always wrongly classified, which suggests that these samples may have very different sequences from the ones used in the training dataset or need to be reclassified.

Table 2. Accuracy results for each genus using the benchmarking dataset, tested with 10 different model ensembles developed here and with VirusTaxo, currently the tool with higher accuracy for virus taxonomy.

Method/ Classification (%)	*Circovirus* (3,647)	*Begomovirus* (118)	*Gemycircularvirus* (42)	*Cyclovirus* (17)	*Porprismacovirus* (31)	*Nanovirus* (10)	*Babuvirus* (125)	*Mastrevirus* (8)	*Gemykibivirus* (18)	*Huchismacovirus* (23)
VirusTaxo	98.43	100.0	80.95	94.11	93.54	90.0	100.0	87.5	94.44	100.0
NoAug	98.71	100.0	80.95	94.11	90.32	90.0	100.0	100.0	94.44	95.65
PercAug	98.16	100.0	80.95	88.23	93.54	90.0	100.0	100.0	94.44	95.65
BordAug	98.62	100.0	80.95	88.23	96.77	90.0	100.0	100.0	94.44	95.65
NoAug +PercAug	98.16	100.0	80.95	88.23	93.54	90.0	100.0	100.0	94.44	95.65
NoAug + BordAug	98.62	100.0	80.95	88.23	96.77	90.0	100.0	100.0	94.44	95.65
NoAug (64)	98.46	100.0	80.95	100.0	93.54	90.0	100.0	100.0	94.44	95.65
PercAug (64)	98.54	100.0	80.95	94.11	93.54	90.0	100.0	100.0	94.44	95.65
BordAug (64)	98.60	100.0	80.95	100.0	93.54	90.0	100.0	100.0	94.44	95.65
NoAug + PercAug (64)	98.68	100.0	80.95	94.11	93.54	90.0	100.0	100.0	94.44	95.65
NoAug + BordAug (64)	98.76	100.0	80.95	100.0	93.54	90.0	100.0	100.0	94.44	95.65

CONCLUSIONS

This work has shown again that CNNs have a really good capacity to differentiate between samples and are able to find conserved patterns in genomic sequences, allowing their use in taxonomic classification. We tested two methodologies for augmentation, where adding fragments of sequences to the borders of each new sequence often yielded better results than with no augmentation. Conversely, adding a percentage of mutation to the sequence yielded worse results in some cases, which could be a consequence of divergent/deleterious mutations made in regions of important patterns during augmentation.

Furthermore, it was possible to confirm that the randomness that exists on the hyperparameters tunning changed the results from one trained model to the other, even maintaining the same architecture, suggesting that creating and comparing models, even using different augmentation techniques, could lead to different, and often better, results. Here we present a novel method for classifying viruses in the phylum *Cressdnaviricota* down to the genus level, achieving results that are better than those obtained with the currently most successful tool, VirusTaxo.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J. and Devin, M., 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv, 1603.04467.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local alignment search tool. J Mol Biol 215, 403-410.

Araújo, F.H., Carneiro, A.C., Silva, R.R., Medeiros, F.N. and Ushizima, D.M., 2017. Redes neurais convolucionais com tensorflow: Teoria e prática. III Escola Regional de Informática do Piauí. Anais - Artigos e Minicursos, pp. 382-406.

Asgari, E. and Mofrad, M.R., 2015. Continuous distributed representation of biological sequences for deep proteomics and genomics. PLOS ONE 10, e0141287.

Auslander, N., Gussow, A.B. and Koonin, E.V., 2021. Incorporating machine learning into established bioinformatics frameworks. Int J Mol Sci 22, 2903.

Baeten, L., Reumers, J., Tur, V., Stricher, F., Lenaerts, T., Serrano, L., Rousseau, F. and Schymkowitz, J., 2008. Reconstruction of protein backbones from the BriX collection of canonical protein fragments. PLOS Comput Biol 4, e1000083.

Barukab, O., Ali, F., Alghamdi, W., Bassam, Y. and Afzal Khan, S., 2022. DBP-CNN: Deep learning-based prediction of DNA-binding proteins by coupling discrete cosine transform with two-dimensional convolutional neural network. Exp Sys Appl 197, 116729.

Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. and de Hoon, M.J., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422-3.

Dasari, C.M. and Bhukya, R., 2022. Explainable deep neural networks for novel viral genome prediction. Appl Intel 52, 3002-3017.

Datta, S., Budhauliya, R., Das, B., Chatterjee, S. and Vanlalhmuaka, V.V., 2015. Next-generation sequencing in clinical virology: Discovery of new viruses. World J Virol 4, 265-276.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv, 1810.04805.

Fabijańska, A. and Grabowski, S., 2019. Viral Genome Deep Classifier. IEEE Access 7, 81297-81307.

Fauquet, C.M. 1999. Taxonomy, classification and nomenclature of viruses. In: Granoff, A. and Webster, R.G. (Eds), Encyclopedia of Virology (2nd Ed.), Elsevier, Oxford, pp. 1730-1756.

Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150-2.

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E.D., Le, Q.V. and Zoph, B. 2021. Simple copy-paste is a strong data augmentation method for instance segmentation, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, pp. 2917-2927.

Gorbalenya, A.E. and Lauber, C., 2022. Bioinformatics of virus taxonomy: foundations and tools for developing sequence-based hierarchical classification. Curr Opin Virol 52, 48-56.

Gorbalenya, A.E., Lauber, C. and Siddell, S. 2019. Taxonomy of Viruses, Reference Module in Biomedical Sciences, Elsevier.

Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. arXiv, 1412.6980.

Koonin, E.V. and Yutin, N., 2020. The crAss-like phage group: How metagenomics reshaped the human virome. Trends Microbiol 28, 349-359.

Krupovic, M., Ghabrial, S.A., Jiang, D. and Varsani, A., 2016. *Genomoviridae*: A new family of widespread single-stranded DNA viruses. Arch Virol 161, 2633-2643.

Krupovic, M., Varsani, A., Kazlauskas, D., Breitbart, M., Delwart, E., Rosario, K., Yutin, N., Wolf, Y.I., Harrach, B. and Zerbini, F.M.J.J.o.v., 2020. *Cressdnaviricota*: A virus phylum unifying seven families of rep-encoding viruses with single-stranded, circular DNA genomes. J Virol 94, e00582-20.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. Nature 521, 436-444.

Li, Z., Liu, F., Yang, W., Peng, S. and Zhou, J., 2022. A survey of convolutional neural networks: Analysis, applications, and prospects. IEEE Trans Neural Net Learn Sys 33, 6999-7019.

Lin, W.-C., Tsai, C.-F., Hu, Y.-H. and Jhang, J.-S., 2017. Clustering-based undersampling in class-imbalanced data. Inf Sci 409-410, 17-26.

Masko, D. and Hensman, P. 2015. The impact of imbalanced training data for convolutional neural networks, Royal Institute of Technology, Stockholm, Sweden.

Min, X., Ye, C., Liu, X. and Zeng, X., 2021. Predicting enhancer-promoter interactions by deep learning and matching heuristic. Brief Bioinform 22, bbaa254.

Mock, F., Kretschmer, F., Kriese, A., Böcker, S. and Marz, M., 2022. Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks. Proc Natl Acad Sci USA 119, e2122636119.

Nair, V. and Hinton, G.E. 2010. Rectified linear units improve restricted boltzmann machines, Proceedings of the 27th International Conference on International Conference on Machine Learning, Omnipress, Haifa, Israel, pp. 807-814.

Ounit, R., Wanamaker, S., Close, T.J. and Lonardi, S., 2015. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics 16, 236.

Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N.N. and Kyrpides, N.C., 2016. Uncovering Earth's virome. Nature 536, 425-30.

Pappagari, R., Villalba, J., Żelasko, P., Moro-Velazquez, L. and Dehak, N., 2021. CopyPaste: An augmentation method for speech emotion recognition. arXiv, 2010.14602.

Pappas, N., Roux, S., Hölzer, M., Lamkiewicz, K., Mock, F., Marz, M. and Dutilh, B.E. 2021. Virus Bioinformatics. In: Bamford, D.H. and Zuckerman, M. (Eds), Encyclopedia of Virology, pp. 124-32.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V., 2011. Scikit-learn: Machine learning in Python. J Mach Learn Res 12, 2825-2830.

Qu, W., Balki, I., Mendez, M., Valen, J., Levman, J. and Tyrrell, P.N., 2020. Assessing and mitigating the effects of class imbalance in machine learning with application to X-ray imaging. Int J Comp Assist Radiol Sur 15, 2041-2048.

Raju, R.S., Al Nahid, A., Chondrow Dev, P. and Islam, R., 2022. VirusTaxo: Taxonomic classification of viruses from the genome sequence using k-mer enrichment. Genomics 114, 110414.

Ren, J., Song, K., Deng, C., Ahlgren, N.A., Fuhrman, J.A., Li, Y., Xie, X., Poplin, R. and Sun, F., 2020. Identifying viruses from metagenomic data using deep learning. Quant Biol 8, 64-77.

Rojas, M.R., Macedo, M.A., Maliano, M.R., Soto-Aguilar, M., Souza, J.O., Briddon, R.W., Kenyon, L.A., Rivera-Bustamante, R.F., Zerbini, F.M., Adkins, S., Legg, J.P., Kvarnheden, A., Wintermantel, W.M., Sudarshana, M.R., Peterschmitt, M., Lapidot, M., Martin, D.P., Moriones, E., Inoue-Nagata, A.K. and Gilbertson, R.L., 2018. World management of geminiviruses. Annu Rev Phytopathol 56, 637-677.

Rosario, K., Duffy, S. and Breitbart, M., 2012. A field guide to eukaryotic circular single-stranded DNA viruses: Insights gained from metagenomics. Arch Virol 157, 1851-1871.

Simmonds, P., Adams, M.J., Benkő, M., Breitbart, M., Brister, J.R., Carstens, E.B., Davison, A.J., Delwart, E., Gorbalenya, A.E., Harrach, B., Hull, R., King, A.M.Q., Koonin, E.V., Krupovic, M., Kuhn, J.H., Lefkowitz, E.J., Nibert, M.L., Orton, R., Roossinck, M.J., Sabanadzovic, S., Sullivan, M.B., Suttle, C.A., Tesh, R.B., van der Vlugt, R.A., Varsani, A. and Zerbini, F.M., 2017. Consensus statement: virus taxonomy in the age of metagenomics. Nat Rev Microbiol 15, 161.

Simmonds, P., Adriaenssens, E.M., Zerbini, F.M., Abrescia, N.G.A., Aiewsakun, P., Alfenas-Zerbini, P., Bao, Y., Barylski, J., Drosten, C., Duffy, S., Duprex, W.P., Dutilh, B.E., Elena, S.F., García, M.L., Junglen, S., Katzourakis, A., Koonin, E.V., Krupovic, M., Kuhn, J.H., Lambert, A.J., Lefkowitz, E.J., Łobocka, M., Lood, C., Mahony, J., Meier-Kolthoff, J.P., Mushegian, A.R., Oksanen, H.M., Poranen, M.M., Reyes-Muñoz, A., Robertson, D.L., Roux, S., Rubino, L., Sabanadzovic, S., Siddell, S., Skern, T., Smith, D.B., Sullivan, M.B., Suzuki, N., Turner, D., Van Doorslaer, K., Vandamme, A.-M., Varsani, A. and Vasilakis, N., 2023. Four principles to establish a universal virus taxonomy. PLOS Biol 21, e3001922.

Wood, D.E., Lu, J. and Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. Genome Biol 20, 257.

Zeng, H. and Gifford, D.K., 2017. Predicting the impact of non-coding variants on DNA methylation. Nucleic Acids Res 45, e99.

Zhang, Y., Qiao, S., Ji, S., Han, N., Liu, D. and Zhou, J., 2019. Identification of DNA–protein binding sites by bootstrap multiple convolutional neural networks on sequence information. Eng Appl Art Intel 79, 58-66.

CHAPTER 2

IN SILICO IDENTIFICATION OF SMALL OPEN READING FRAMES (sORFs) IN ALPHASATELLITE GENOMES

Gomes RAL, Zerbini FM. *In silico* identification of small open reading frames (sORFs) in alphasatellite genomes.

In silico identification of small open reading frames (sORFs) in alphasatellite genomes

Ruither A. L. Gomes[1,2], F. Murilo Zerbini[1,2*]

[1]Dep. de Fitopatologia, [2]National Institute for Science and Technology on Plant-Pest Interactions, Universidade Federal de Viçosa, Viçosa, MG, 36570-900, Brazil;

*Corresponding author:
Phone: (+55-31) 2612-2423; E-mail: zerbini@ufv.br

ABSTRACT

Among the various innovations in biochemistry and molecular biology in the past decades, high-throughput DNA sequencing and mass spectrometry have revigorated the study of proteomics, particularly the discovery and functional characterization of proteins encoded from small open reading frames (sORFs). Viruses in general, and plant viruses in particular, have very small genomes that are highly coding-dense. Thus, the presence of sORFs is probably more widespread in viral genomes than what is currently known. A recent study demonstrated the functional importance of sORFs in with the genome of tomato yellow leaf curl virus (TYLCV), a member of the genus *Begomovirus*. Begomoviruses are frequently associated with different types of DNA satellites. One of them, the alphasatellites, recently have been shown to increase disease symptoms and viral DNA accumulation when associated with two New World (NW) begomoviruses, while those associated with Old World (OW) begomoviruses do not cause these effects. The objective of this work was to detect the presence of functional sORFs in NW and OW alphasatellites. Five NW alphasatellites were shown to have two conserved sORFs with predicted functional domains. These sORFs were not detected in the six OW alphasatellites that were tested. One of the OW alphasatellites has an sORF which is similar to one found in NW alphasatellites, but with no functional domain. Thus, it is possible that the sORFs present in NW alphasatellites play a role in the infection cycle. Further studies are necessary to confirm this hypothesis.

INTRODUCTION

Great innovations in technology, mainly in the last few dozen years, have increased considerably our capacity to study science in all biological fields. These innovations include a wide range of developments in machine learning and computational biology, and also in high-throughput analytical techniques, such as mass spectrometry (MS) and DNA sequencing. Specifically for virology, these improvements are helping us to study the virosphere's stunning (and somehow worrying) diversity, e.g. with computational methods to search for viral sequences, extract their functional information, and organize the acquired knowledge in an understandable way (Pappas et al., 2021). However, even with the advent of high-throughput DNA sequencing (HTS), the correct annotation of all generated information is still a challenging and crucial step in the path to better understanding viruses (Finkel et al., 2018).

An open reading frames (ORFs) is defined as an in-frame DNA sequence of codons (three consecutive nucleotides) between a start codon and a stop codon. This structural conservation made ORF detection the purpose of some of the first computational tools developed for DNA sequence analysis. In general, these tools default parameters search for ORFs encoding at least 50 amino acids (aa), in consequence of their higher statistical confidence (the odds of a random, non-functional DNA sequence having the features of an ORF are much lower the longer its length). Very few functional ORFs were found below this threshold until recently (Basrai et al., 1997).

Once genomes became easily accessible by HTS, the development of better annotation tools led to an increasing number of functional short ORFs (sORFs) being identified in different studies (Schlesinger and Elsässer, 2022). These sORFs were shown to have important roles in different organisms throughout the tree of life, and in the last decade the search for, and the characterization of, sORFs became a focus of research on many different groups of organisms, as well as on the development of ever more sophisticated algorithms for the search for these

sORFs (Albuquerque et al., 2015; Andrews and Rothnagel, 2014; Guerra-Almeida et al., 2021; Hellens et al., 2016; Hsu and Benfey, 2018; Makarewich and Olson, 2017; Peeters and Menschaert, 2020; Ruiz-Orera and Albà, 2019; Schlesinger and Elsässer, 2022).

sORFs detected by bioinformatics-based approaches must be functionally validated. The most common high-throughput tools for validation include ribosome profiling and mass spectrometry. Ribosome profiling provides evidence that sORFs are translated by sequencing RNA fragments that were protected from nucleases by being "inside" ribosomes, while mass spectrometry is used to detect the peptide(s) translated from sORFs (often referred to as micropeptides). From a computational side, the differentiation of truly expressed proteins from random sequences with no function can be done by analyzing the sequence metrics of conservation or similarity with other known sequences, with the aid of different tools (Peeters and Menschaert, 2020).

When searching for conservation, one of the commonly used approaches for ORFs in general is to check if the sequence shows more synonymous mutations than non-synonymous ones, which is a good indication that no meaningful changes occur in the encoded protein. However, this technique does not work so well for short sequences, as they tend to produce less reliable substitution rates (Schlesinger and Elsässer, 2022). In this context, specific tools were developed for the detection and analysis of sORFs, such as sORF finder (Hanada et al., 2009), a tool that uses nucleotide composition bias and synonymous/nonsynonymous substitution rates but that requires a considerable amount of coding and noncoding sequences to be available in the analyzed genome, having been tested using *Saccharomyces cerevisiae* and *Arabidopsis thaliana* genomes.

For annotation of genomic sequences based on similarity to a previously identified group of proteins of interest, well-known tools such as BLAST (Altschul et al., 1990) or HMMER (Eddy, 1995) are still commonly applied. These tools point to possible similar

functions between the sORFs detected in the genome and previously characterized (and thus functionally active) ORFs (Schlesinger and Elsässer, 2022). Moreover, is possible to use the deduced amino acid sequence directly for functional annotation by comparison with curated protein databases or by applying different protein function prediction tools, which nowadays can both be done in one place with the recently released version of the InterPro website (ebi.ac.uk/interpro/), currently the "world's most comprehensive resource about proteins families, domains and functional sites" (Paysan-Lafosse et al., 2023).

Despite the existence of "giant" viruses with genomes with more than 1 million nucleotides (nt) (Koonin and Yutin, 2018), the vast majority of viruses have small genomes (<50,000 nt). Plant viruses in particular have very small genomes, ranging from 2,600 to 20,000 nt (Hull, 2014). Thus, viral genomes are usually highly coding-dense, often with overlapping ORFs in both senses. It is therefore not surprising that viruses also tend to have sORFs that encode small functional proteins, something that was recently demonstrated for the plant-infecting geminiviruses (whose genomes are among the smallest of all viruses, as small as 2,600 nt) (Gong et al., 2021). Moreover, as reviewed by Finkel et al. (2018), one consequence of viruses' fast evolution rates is the increased probability of protein function preservation when the sequence is conserved among genomes. They also reviewed findings of functional viral sORFs, and also showed that these sequences tend to be enriched for specific functional features in some virus groups.

The study of Gong et al. (2021), mentioned above, used NCBI's ORFfinder tool (ncbi.nlm.nih.gov/orffinder/) together with a custom-built tool to find viral sORFs, and BLASTp for grouping the most similar proteins. They were able to find six sORFs with functional potential in the genome of tomato yellow leaf curl virus (TYLCV), and functional studies indicated that at least one of the proteins encoded by these sORFs is required for

systemic viral infection and was localized to the Golgi apparatus, where it acts as an RNA silencing suppressor.

TYLCV belongs to the family *Geminiviridae*, which includes viruses with circular, single-stranded (ss) DNA genomes that infect plants. The family has 14 genera, and viruses in the genus *Begomovirus* (which includes TYLCV) may have monopartite or bipartite genomes, with each genomic component ranging in length from 2,500 to 2,700 nt (Fiallo-Olive et al., 2021).

Among the viruses in the family, begomoviruses have the spotlight as they cause greater losses in economically important dicotyledonous crop plants in all tropical and subtropical regions of the world (Rojas et al., 2018). The genus *Begomovirus* comprises the majority of the species in the family (445 out of 520). Begomoviruses are transmitted by a complex of cryptic species of whiteflies (*Bemisia tabaci*) and can be divided into two groups based on genome architecture and phylogenetics, New World (NW), with mostly bipartite viruses rarely associated with DNA satellites, and Old World (OW), with mostly monopartite viruses often associated with DNA satellites (Fiallo-Olive et al., 2021; Zhou, 2013).

Begomovirus-associated DNA satellites are circular, ssDNA molecules with half of the length of begomovirus genomes (approx. 1,300 nt) which are dependent on a helper begomovirus for one or more essential functions (replication, movement or encapsidation) (Zhou, 2013). The associated begomovirus provides proteins to carry out these functions during infection. There are three types of DNA satellites, alpha-, beta- and deltasatellites (Briddon et al., 2003; Lozano et al., 2016; Saunders and Stanley, 1999). The genomes of alphasatellites contain in the standard stem-loop structure found in begomoviruses (which includes the origin of replication), an adenine-rich region, and a single ORF that encodes a protein named alpha-Rep, which is associated with replication (Briddon et al., 2018; Nogueira et al., 2021; Zhou,

2013). Alphasatellites are also associated with viruses in another family of circular, ssDNA viruses, the *Nanoviridae* (Briddon et al., 2018).

Recent studies with NW alphasatellites have shown that, in contrast with the lack of an effect or attenuation of disease symptoms seen in infections with some OW alphasatellites (Idris et al., 2011), NW ones cause an increase in symptoms, with an impact on the helper virus accumulation depending on the combination of host, alphasatellite and helper virus (Mar et al., 2015; Nogueira et al., 2021). However, is not yet clear how alphasatellites interact with their helper viruses. Seeking to better elucidate these interactions and what could be leading to this increase in symptoms, the objective of this work was to use computational approaches to investigate the presence of functional sORFs in the genomic sequences of alphasatellites.

MATERIALS AND METHODS

Data acquisition

Two of the NW alphasatellites analyzed here, Euphorbia yellow mosaic alphasatellite (EuYMA) and tomato yellow spot alphasatellite (ToYSA), were shown by Nogueira et al. (2021) to cause an increase in the severity of symptoms during infection with helper begomoviruses. Three additional NW alphasatellites were also analyzed: croton yellow vein mosaic alphasatellite (CrYVMA), melon chlorotic mosaic alphasatellite (MeCMA), and whitefly-associated Puerto Rico alphasatellite 1 (WfaPRA1). The six OW alphasatellites analyzed, which were shown previously not to have an effect, or attenuate the symptoms of infection, were Ageratum enation alphasatellite (AEA), Ageratum yellow vein alphasatellite (AYVA), Ageratum yellow vein Singapore alphasatellite (AYVSA), cotton leaf curl Multan alphasatellite (CLCuMuA), cotton leaf curl Gezira alphasatellite (CLCuGeA) and tomato yellow leaf curl China alphasatellite (TYLCCNA) (Luo et al., 2019; Saunders et al., 2000).

All isolates of the selected alphasatellites used in this work were downloaded from the NCBI GenBank database (ncbi.nlm.nih.gov/genbank) using the accession numbers listed by Briddon et al. (2018). All sequences are available in the *Input* folder at the same link as the supplementary files.

Identification and functional prediction of sORFs in alphasatellite genomes

Based on the idea that "sequence conservation is highly indicative of conserved function" (Finkel et al., 2018), the presence of functional sORFs was assessed by comparisons of translated sORFs sequences among different species.

For that, a custom Python script was built (and is available at the same link as the supplementary files) to initially translate all six reading frames, saving only protein sequences with more than six amino acids (from the first methionine to the stop codon) from each genome,

then using BLASTp (Altschul et al., 1990) for clustering similar proteins (using identity and coverage higher than 50% as a clustering threshold filter) and generating as output a size-ordered list of all groups of sORFs identified for a given species.

For assessing the existence of functional domains, all sORFs identified as conserved were manually tested on the InterPro website (ebi.ac.uk/interpro/), using the freely available integrated prediction tool InterProScan. Also, deduced amino acid sequences that exhibited predicted domains/regions were checked for similarity with other proteins using the UniProt web BLAST (uniprot.org/blast), using the standard parameters and database. Additional data analyses, such as sequence alignments or phylogenetic trees, were done using MEGA11 (Tamura et al., 2021) using standard parameters.

RESULTS AND DISCUSSION

The initial dataset was composed of 12 EuYMA, 5 ToYSA, 4 AYVA, 2 AYVSGA, and 4 TYLCCNA sequences. Sequences were parsed by our custom script generating two output files of size-ordered proteins found on each genome, one with all ORFs detected and one only with ORFs that were present in all input sequences (conserved sORF files are available in the *Output* folder at the same link as the supplementary files).

For EuYMA, eight sORFs were found to be conserved in all 12 isolates (Suppl. file 1), and from these, the protein products of three ORFs showed the presence of domains according to InterProScan (Figures 1-3; all output files from this tool are available in a specific folder within the supplementary files link), using KY559640 as the representative sequence.

The first sORF whose protein product displays a predicted domain (henceforth called sORF1) is longer than normally expected from sORFs (441 nt, potentially encoding a 147 aa protein) but, to the best of our knowledge, is uncharacterized in EuYMA, so it was treated as a sORF for the objective of this work. The deduced aa sequence was predicted to possess an intrinsically disordered region (IDR) by the MobiDB-lite predictor (Piovesan et al., 2020; Figure 1), which in viruses is a region generally described as acting on reprogramming host cells, with activities related to signal transduction, interaction with proteins, regulation of gene expression, and others (Davey et al., 2011; Xue et al., 2010). When assessing for the existence of similarity with other classified proteins using UniProt BLAST (also with the output files available in a specific folder within the supplementary files link), an identity of 76.3% and *E*-value of 10^{-68} was found with a putative protein named alpha-V2 from MeCMA. Interestingly, MeCMA is another NW alphasatellite which is phylogenetically close to EuYMA (both agents are classified in the same genus, *Clecrusatellite*, of the family *Alphasatellitidae*; Briddon et al., 2018).

The protein products of the second and third sORFs (sORF2 and sORF3) were also predicted as possessing IDRs. sORF2 has 168 nt and the potential to encode a 56 aa protein, and the IDR encompasses aa positions 1 to 31 (Figure 2). This protein's BLAST analysis indicated a 44% identity, with an *e*-value of $7.9x10^{-4}$, with a small portion of a larger protein related to lipid transporter and nutrient reservoir named Vitellogenin 2. Besides the low similarity and high *E* value, vitellogenins are found only in birds, which are not known hosts for nanovirids, from which alphasatellites are thought to have originated (Briddon et al., 2018). The sORF3 has 84 nt, and the potential to encode a 28 aa protein, and the IDR encompasses aa positions 1 to 28 (Figure 3). BLAST analysis did not indicate any identity with other proteins, with the top hit displaying an *E*-value of 0.58.

Figure 1. Analysis of the deduced amino acid sequence of Euphorbia yellow mosaic alphasatellite (EuYMA) sORF1 (441 nucleotides). An intrinsically disordered region was predicted by MobiDB-lite v. 1, encompassing amino acid positions 120 to 147.

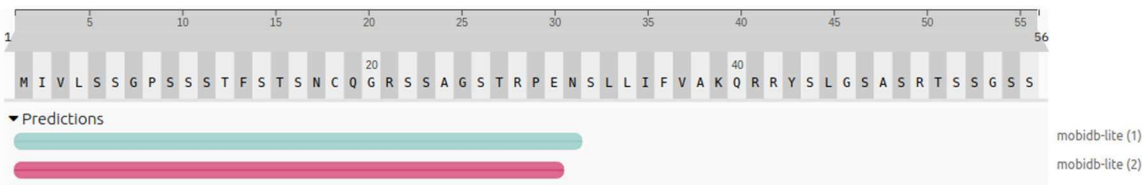

Figure 2. Analysis of the deduced amino acid sequence of Euphorbia yellow mosaic alphasatellite (EuYMA) sORF2 (168 nucleotides). An intrinsically disordered region was predicted by MobiDB-lite v. 1 and 2, encompassing amino acid positions 1 to 30 (v. 1) or 31 (v. 2).

Figure 3. Analysis of the deduced amino acid sequence of Euphorbia yellow mosaic alphasatellite (EuYMA) sORF3 (84 nucleotides). An intrinsically disordered region was predicted by MobiDB-lite v. 1 and 2, encompassing amino acid positions 1 (v. 1) or 12 (v. 2) to 28.

The second NW alphasatellite evaluated was ToYSA, in which 23 sORFs were detected that were conserved in all five sequenced isolates (Suppl. file 2). This high number is likely due to the small number of isolates available, increasing the probability that several of these sORFs occur by chance. Nevertheless, it was possible to find predicted domains in the protein products of the four larger sORFs (Figures 4-7) using KX348228 as the representative sequence.

The first sORF identified is similar to sORF1 from EuYMA, with approximately the same length (429 nt), and the potentially encoded 143 aa protein) has a predicted IDR at almost the same position (Figure 4), also displaying high identity (77.7%) and a low E-value ($1.6x10^{-69}$) with alpha-V2 from MeCMA. These results are expected since ToYSA is closely phylogenetically related to EuYMA (Ferro et al., 2017).

The second sORF identified is also similar to sORF2 in EuYMA. Again, both have the same length (168 nt, potentially encoding a 56 aa protein), with an IDR spanning aa positions 1 to 31 (Figure 5) and 46.2% identity (E-value of $3x10^{-3}$) with vitellogenin 2.

The sORF4 (81 nt, potentially encoding a 27 aa protein) has no homolog in EuYMA. The TMHMM tool (Krogh et al., 2001) predicted a transmembrane domain in the protein spanning aa positions 7 to 26 (Figure 6). Blast analysis indicated a degree of similarity (58.2% identity, *E*-value of $9x10^{-3}$) with membrane proteins mainly from one cuttlefish species, which

again considering the sORF size and the small number of isolates, could be a product of randomness.

The shortest ToYSA sORF analyzed was sORF5, which is 72 nt long and potentially encodes a 24 aa protein. The entire length of the protein is predicted to be an IDR (Figure 7). This protein had no significant hits with other proteins based on UniProt BLAST.

Considering that only sORF1 and sORF2 were present in both EuYMA and ToYSA, these two sORFs were the ones used for comparison with OW alphasatellites (whose presence, unlike that of their NW counterparts, does not increase disease symptoms or helper virus load).

Figure 4. Analysis of the deduced amino acid sequence of tomato yellow spot alphasatellite (ToYSA) sORF1 (429 nucleotides). An intrinsically disordered region was predicted by MobiDB-lite, encompassing amino acid positions 122 to 147.

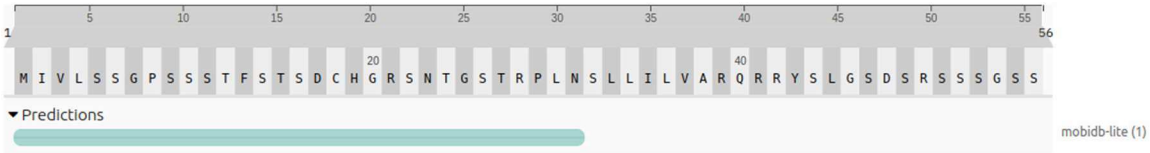

Figure 5. Analysis of the deduced amino acid sequence of tomato yellow spot alphasatellite (ToYSA) sORF2 (168 nucleotides). An intrinsically disordered region was predicted by MobiDB-lite, encompassing amino acid positions 1 to 31.

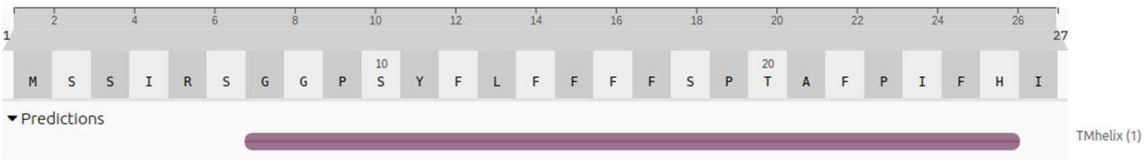

Figure 6. Analysis of the deduced amino acid sequence of tomato yellow spot alphasatellite (ToYSA) sORF4 (81 nucleotides). Model TMhelix v. 1 predicted a transmembrane region encompassing amino acid positions 7 to 26.

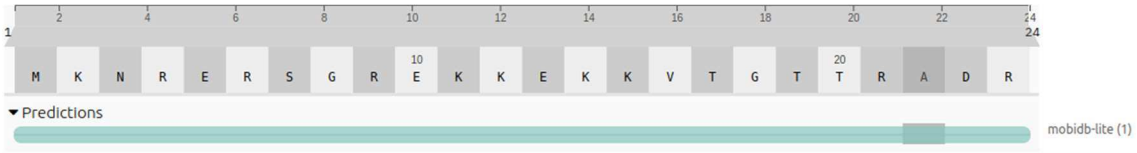

Figure 7. Analysis of the deduced amino acid sequence of tomato yellow spot alphasatellite (ToYSA) sORF5 (72 nucleotides). An intrinsically disordered region was predicted by MobiDB-lite, encompassing amino acid positions 1 to 24.

The first OW alphasatellite analyzed was Ageratum yellow vein alphasatellite (AYVA), with four isolates. This alphasatellite has five sORFs which are conserved in all four isolates (Suppl. file 3), but none of them are similar to EuYMA/ToYSA sORF1 or sORF2. When checking with InterProScan, a sORF (sORF6, with 135 nt) in one of the isolates potentially encoded a protein with a region identified as related to the superfamily PIRSF029928 (UniRule UR000764442; MacDougall et al., 2020) (Figure 8). This is a group of bacterial proteins with functions associated with transformation, DNA binding, methylation, competence, transport, and others. However, this sORF was 15 nt shorter in the other three AYVA isolates, and the absence of the last 5 aa in the deduced protein sequence caused the region not to be predicted anymore. Thus, this putative protein was not further analyzed.

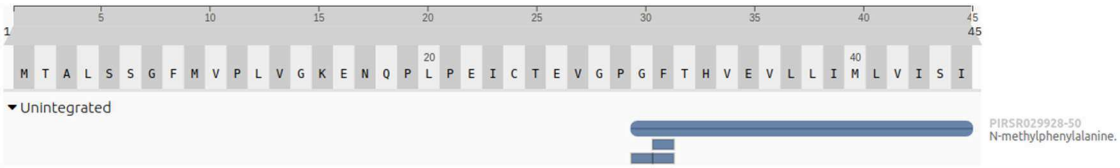

Figure 8. Analysis of the deduced amino acid sequence of Ageratum yellow vein alphasatellite (AYVA) sORF6 (135 nucleotides). A domain related to the superfamily PIRSF029928 was predicted by InterProScan, encompassing amino acid positions 30 to 45.

Only two isolates of AYVSGA have been sequenced, and therefore it was not surprising that 13 sORFs were "conserved" in both of them (Supplementary file 4). However, none of the proteins potentially encoded by these sORFs had any predicted domains or regions. The longest sORF (486 nt, potentially encoding a protein with 162 aa) is in the same relative position in the genome and has considerable sequence identity with, sORF1 from EuYMA/ToYSA, but the protein has no predicted domain. A second sORF (213 nt, potentially encoding a 71 aa protein) is in the same relative position as EuYMA/ToYSA sORF2, but the protein also has no predicted domains. Interestingly, these similarities are consistent with the close evolutionary relationship between this OW alphasatellite and the NW ones (Briddon et al., 2018; Mar et al., 2015).

The third OW alphasatellite to be analyzed was TYLCCNA, with four sequenced isolates. A total of 10 sORF were found to be conserved among the four isolates (Suppl. file 5). From these, sORF3 (81 nt) potentially encodes a protein with a predicted transmembrane domain (Figure 9), similar to the one detected in the ToYSA sORF4 product, albeit with a different sequence. Again, considering the short sequence and the small number of isolates, this protein was not further analyzed.

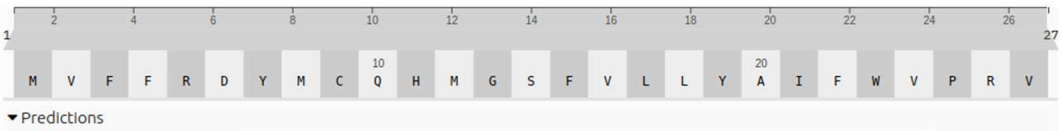

Figure 9. Analysis of the deduced amino acid sequence of tomato yellow leaf curl China alphasatellite (TYLCNCA) sORF3. Model TMhelix version 1 predicted a transmembrane region encompassing amino acid positions 7 to 24.

While these results indicate that the two sORFs which are conserved in the two NW alphasatellites analyzed (EuYMA and ToYSA) were not present, either in full length or with the same predicted domains, in the OW alphasatellites, the small number of isolates available prevents a more definitive conclusion. Thus, as an attempt to enhance the confidence of the results, we tested additional alphasatellites for which a larger number of isolates are available: the NW croton yellow vein mosaic alphasatellite (CrYVMA) with 11 isolates, melon chlorotic mosaic alphasatellite (MeCMA) with 37 isolates, and whitefly-associated Puerto Rico alphasatellite 1 (WfaPRA-1) with 3 isolates, and the OW Ageratum enation alphasatellite (AEA) with 166 isolates, cotton leaf curl Multan alphasatellite (CLCuMuA) with 161 isolates, and cotton leaf curl Gezira alphasatellite (CLCuGeA) with 11 isolates. The three NW alphasatellites are members of the genera *Clecrusatellite*, while the three OW alphasatellites are all members of the genus *Colecusatellite* (Briddon et al., 2018).

Starting with the NW MeCMA, seven sORFs were found to be conserved in all 37 isolates (Suppl. file 6). The protein products of the five shortest sORFs had no predicted conserved domains or regions. The longest sORF1 has different length variants amongst the isolates: 474 nt (158 deduced aa; three isolates), 441 (147 deduced aa; four isolates) and 417 nt (139 deduced aa; the remaining 30 isolates). No conserved domains were detected in the 139 and 147 aa proteins, but an IDR was predicted for the 158 aa protein (Figure 10). sORF2 has 126 nt, but its 42 aa protein product had no predicted domains.

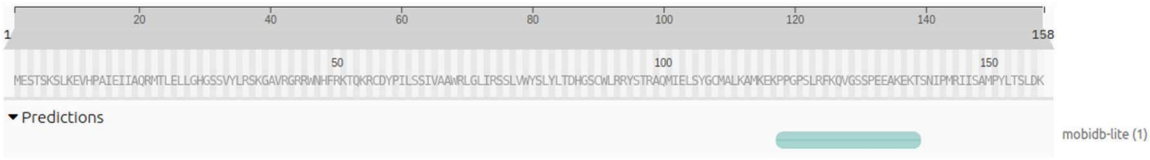

Figure 10. Analysis of the deduced amino acid sequence of melon chlorotic mosaic alphasatellite (MeCMA) sORF1. An intrinsically disordered region was predicted by MobiDB-lite, encompassing amino acid positions 118 to 139.

To better understand these variants, an alignment (using MEGA) of the deduced aa sequences of sORF1 from EuYMA, ToYSA, and the three MeCMA variants was obtained (available on a specific folder within the supplementary files link). From the alignment, it was possible to identify mutations in the region encompassing the IDR domain which could explain the absence of the IDR domain in the two shorter variants. An alignment for sORF2 was also obtained, allowing the identification of three mutations in the IDR domain of MeCMA that could be responsible for the domain no longer being detected. The absence of predicted domains in most MeCMA isolates is interesting as it seems to match the results from Romay *et al.* (2015), who used MeCMA with its helper virus (melon chlorotic mosaic virus, MeCMV) to assess the infection effect of the alphasatellite and concluded that MeCMA didn't have an obvious effect on symptoms. Thus, it is not unreasonable to speculate that these sORFs are not only expressed during infection but also that the presence of the IDR domain in their encoded proteins may be associated with the increase in disease symptoms observed when EuYMA and ToYSA are present.

When analyzing CrYVMA, we noticed that one isolate (HQ668024) did not have the alpha-Rep ORF. Considering that this ORF is absolutely conserved in all alphasatellites, this isolate was removed from the dataset. The remaining 10 isolates had seven conserved sORFs

(Suppl. file 7). As observed for MeCMA, the product of sORF1 (417 nt, potentially encoding a 139 aa protein) had a predicted IDR domain in its C-terminal portion (Figure 11). A full-length homolog of MeCMA sORF2 was not present, with only a truncated sORF (39 nt and no predicted domains).

Figure 11. Analysis of the deduced amino acid sequence of croton yellow vein mosaic alphasatellite (CrYVMA) sORF1. An intrinsically disordered region was predicted by MobiDB-lite v. 2, encompassing amino acid positions 119 to 139.

The last NW species tested was WfaPRA-1. A total of 17 sORFs were found to be conserved in the three isolates (Suppl. file 8). The longest sORF, named sORF1, was 417 nt long (potentially encoding a 139 aa protein). No conserved domains were predicted for the encoded protein. The aa sequence alignment of the WfaPR-1 sORF1 product with those from EuYMA and ToYSA shows that the WfaPRA-1 sORF1 product has a considerable number of mutations in its C-terminal portion, which is likely the reason why an IDR was not detected as it was in the proteins encoded by the other two alphasatellites.

On the other hand, the product of sORF2 (126 nt, potentially encoding a 42 aa protein) had a predicted IDR (Figure 12) extending through its entire sequence. The aa sequence alignment of the CrYVMA, MeCMA and WfaPRA-1 sORF2 products indicated that the size variation occurred at the C-terminal portion of the protein.

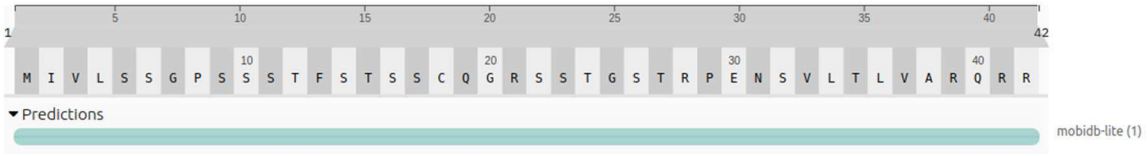

Figure 12. Analysis of the deduced amino acid sequence of whitefly-associated Puerto Rico alphasatellite 1 (WfaPRA-1 sORFs. An intrinsically disordered region was predicted by MobiDB-lite v. 1, encompassing amino acid positions 1 to 42.

Finally, the last three OW alphasatellites were tested, first parsing AEA with, 166 isolates. Interestingly, not a single sORF was found to be conserved in all isolates (one isolate, HQ631431, did not have an alpha-Rep ORF and was removed from the dataset). One sORF (120 or 135 nt, depending on the isolate) had high similarity with the AYVA sORF6, and the protein product had different predicted domains for each of the two length variants (Figures 13 and 14).

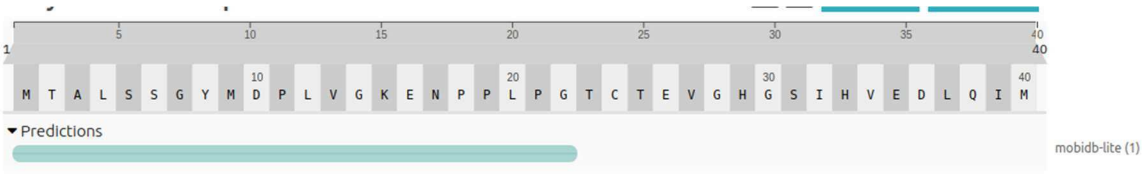

Figure 13. Analysis of the deduced amino acid sequence (40 aa) of Ageratum enation alphasatellite (AEA) sORF6. An intrinsically disordered region was predicted by MobiDB-lite v. 1, encompassing amino acid positions 1 to 22.

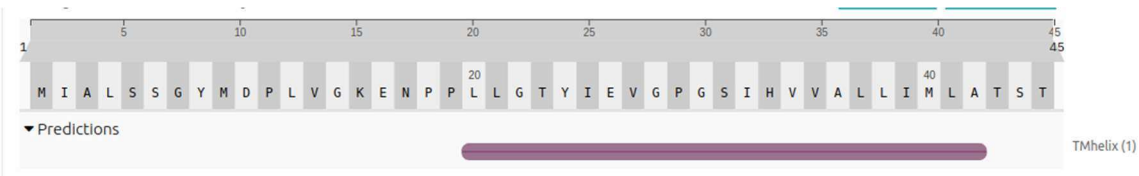

Figure 14. Analysis of the deduced amino acid sequence (45 aa) of Ageratum enation alphasatellite (AEA) sORF6. Model TMhelix v. 1 predicted a transmembrane region encompassing amino acid positions 20 to 42.

The second OW alphasatellite tested was CLCuMuA, with 161 isolates. One sORF was found to be conserved in all isolates (Suppl. file 9), again with two variants of 120 and 135 nt. This sORF was also similar to the AYVA sORF6, but the protein products had no predicted domains.

The last alphasatellite tested was CLCuGeA, with 11 isolates. Two sORF were conserved in all isolates (Suppl. file 10). The product of the longest sORF (156 nt, 52 aa) had no predicted domain. The second sORF was again similar to sORF6 from AYVA, with protein length being either 37 or 45 aa. The 45 aa protein, the most frequent one, and was predicted to have an IDR (Figure 151). The lack of any sORFs similar to sORF1 or sORF2 in OW species suggests that both sORFs are probably encoded only by NW alphasatellites. These results also suggest that sORF6 could be functional in some OW alphasatellites.

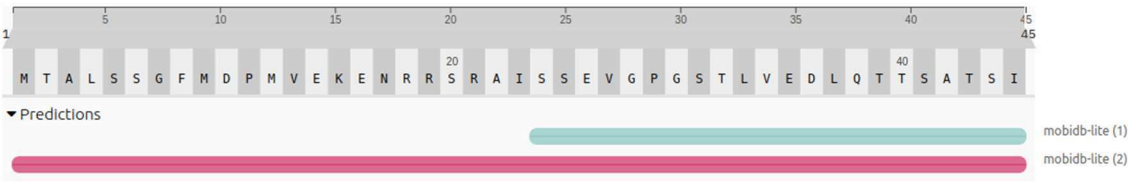

Figure 15. Analysis of the deduced amino acid sequence (45 aa) of cotton leaf curl Gezira alphasatellite (CLCuGeA) sORF. An intrinsically disordered region was predicted by MobiDB-lite v. 2, encompassing amino acid positions 1 to 45.

For better visualization of the two main ORFs found (named 1 and 2), we used the software APE (Davis and Jorgensen, 2022) to generate the genomic map for the alphasatellite with the two known regions and them (Figure 16). It is possible to notice that both overlap the Rep gene, but in a different reading frame (sORF1) and in the anti-sense strand (sORF 2).

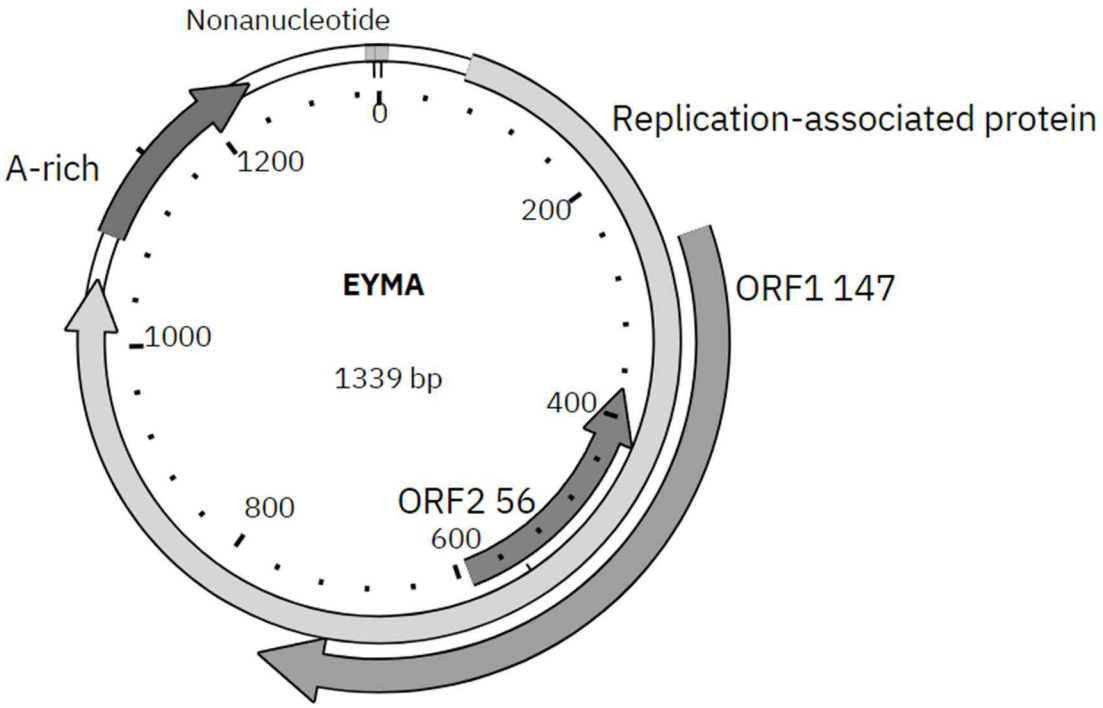

Figure 16. Genomic map of Euphorbia yellow mosaic alphasatellite (EuYMA). The two short ORFs identified in this study are indicated as ORF1 (147 amino acids) and ORF2 (56 amino acids).

CONCLUSIONS

The viral proteome is vastly undersampled. Recent technologies have opened new avenues of study on how different viruses achieve all necessary functions for their replication and spread. Viruses are fast-evolving entities that need to compress genomic information in small genomes, being ideal targets for studying the functions and the relevance of sORFs.

In this work we searched for the presence of functional sORFs in different alphasatellites, seeking those that could be related to the differences in the effect on disease symptoms that have been reported between NW and OW alphasatellites. We were able to identify two NW sORFs that were present in most NW alphasatellites with predicted domains related to intrinsically disordered proteins. We also identified a sORF from OW alphasatellites that haves predicted domains in some isolates. Further studies are necessary to demonstrate that these sORFs are indeed functional, and to determine their role in the infection cycle.

SUPPLEMENTARY INFORMATION

All supplementary information is available in a Google Drive folder at the link: https://drive.google.com/drive/folders/1zlnX6wbMlogjPtszB3WAq_Eh0RhXWKyG?usp=sharing. All json files are results from the InterProScan, and can be downloaded and imported for visualization on the results web page, available at https://www.ebi.ac.uk/interpro/result/InterProScan/#table.

REFERENCES

Albuquerque, J.P., Tobias-Santos, V., Rodrigues, A.C., Mury, F.B., Fonseca, R.N.d., 2015. v. Genet Mol Biol 38, 278-283.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J Mol Biol 215, 403-410.

Andrews, S.J., Rothnagel, J.A., 2014. Emerging evidence for functional peptides encoded by short open reading frames. Nat Rev Genet 15, 193-204.

Basrai, M.A., Hieter, P., Boeke, J.D., 1997. Small open reading frames: Beautiful needles in the haystack. Genome Res 7, 768-771.

Briddon, R.W., Bull, S.E., Amin, I., Idris, A.M., Mansoor, S., Bedford, I.D., Dhawan, P., Rishi, N., Siwatch, S.S., Abdel-Salam, A.M., Brown, J.K., Zafar, Y., Markham, P.G., 2003. Diversity of DNA beta, a satellite molecule associated with some monopartite begomoviruses. Virology 312, 106-121.

Briddon, R.W., Martin, D.P., Roumagnac, P., Navas-Castillo, J., Fiallo-Olive, E., Moriones, E., Lett, J.M., Zerbini, F.M., Varsani, A., 2018. *Alphasatellitidae*: A new family with two subfamilies for the classification of geminivirus- and nanovirus-associated alphasatellites. Arch Virol 163, 2587-2600.

Davey, N.E., Travé, G., Gibson, T.J., 2011. How viruses hijack cell regulation. Trends Biochem Sci 36, 159-169.

Davis, M.W., Jorgensen, E.M., 2022. ApE, A Plasmid Editor: A Freely Available DNA Manipulation and Visualization Program. Front Bioinf 2, 818619.

Eddy, S.R., 1995. Multiple alignment using hidden Markov models. Proc Int Conf Intell Syst Mol Biol 3, 114-120.

Ferro, C.G., Silva, J.P., Xavier, C.A.D., Godinho, M.T., Lima, A.T.M., Mar, T.B., Lau, D., Zerbini, F.M., 2017. The ever increasing diversity of begomoviruses infecting non-cultivated hosts: new species from *Sida* spp. and *Leonurus sibiricus*, plus two New World alphasatellites. Ann Appl Biol 170, 204-218.

Fiallo-Olive, E., Lett, J.M., Martin, D.P., Roumagnac, P., Varsani, A., Zerbini, F.M., Navas-Castillo, J., 2021. ICTV Virus Taxonomy Profile: *Geminiviridae* 2021. J Gen Virol 102, 001696.

Finkel, Y., Stern-Ginossar, N., Schwartz, M., 2018. Viral short ORFs and their possible functions. Proteomics 18, e1700255.

Gong, P., Tan, H., Zhao, S., Li, H., Liu, H., Ma, Y., Zhang, X., Rong, J., Fu, X., Lozano-Duran, R., Li, F., Zhou, X., 2021. Geminiviruses encode additional small proteins with specific subcellular localizations and virulence function. Nat Commun 12, 4278.

Guerra-Almeida, D., Tschoeke, D.A., Nunes-da-Fonseca, R., 2021. Understanding small ORF diversity through a comprehensive transcription feature classification. DNA Res 28, dsab007.

Hanada, K., Akiyama, K., Sakurai, T., Toyoda, T., Shinozaki, K., Shiu, S.-H., 2009. sORF finder: a program package to identify small open reading frames with high coding potential. Bioinformatics 26, 399-400.

Hellens, R.P., Brown, C.M., Chisnall, M.A.W., Waterhouse, P.M., Macknight, R.C., 2016. The emerging world of small ORFs. Trends Plant Sci 21, 317-328.

Hsu, P.Y., Benfey, P.N., 2018. Small but mighty: Functional peptides encoded by small ORFs in plants. Proteomics 18, 1700038.

Hull, R., 2014. Plant Virology, 5th Ed. ed. Elsevier Academic Press, San Diego, USA.

Idris, A.M., Shahid, M.S., Briddon, R.W., Khan, A.J., Zhu, J.K., Brown, J.K., 2011. An unusual alphasatellite associated with monopartite begomoviruses attenuates symptoms and reduces betasatellite accumulation. J Gen Virol 92, 706-717.

Koonin, E.V., Yutin, N., 2018. Multiple evolutionary origins of giant viruses. F1000 Res 7, 1840.

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L.L., 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. J Mol Biol 305, 567-580.

Lozano, G., Trenado, H.P., Fiallo-Olive, E., Chirinos, D., Geraud-Pouey, F., Briddon, R.W., Navas-Castillo, J., 2016. Characterization of non-coding DNA satellites associated with sweepoviruses (genus *Begomovirus*, *Geminiviridae*) - definition of a distinct class of begomovirus-associated satellites. Front Microbiol 7, 162.

Luo, C., Wang, Z.Q., Liu, X., Zhao, L., Zhou, X., Xie, Y., 2019. Identification and analysis of potential genes regulated by an alphasatellite (TYLCCNA) that contribute to host resistance against tomato yellow leaf curl China Virus and its betasatellite (TYLCCNV/TYLCCNB) infection in *Nicotiana benthamiana*. Viruses 11, 442.

MacDougall, A., Volynkin, V., Saidi, R., Poggioli, D., Zellner, H., Hatton-Ellis, E., Joshi, V., O'Donovan, C., Orchard, S., Auchincloss, A.H., Baratin, D., Bolleman, J., Coudert, E., de Castro, E., Hulo, C., Masson, P., Pedruzzi, I., Rivoire, C., Arighi, C., Wang, Q., Chen, C., Huang, H., Garavelli, J., Vinayaka, C.R., Yeh, L.-S., Natale, D.A., Laiho, K., Martin, M.-J., Renaux, A., Pichler, K., Consortium, T.U., 2020. UniRule: a unified rule resource for automatic annotation in the UniProt Knowledgebase. Bioinformatics 36, 4643-4648.

Makarewich, C.A., Olson, E.N., 2017. Mining for micropeptides. Trends Cell Biol 27, 685-696.

Mar, T.B., Alves, M.S., Barbosa, L.R., Amaral, J.G., Pereira, H.M.B., Mendes, I.R., Fiallo-Olive, E., Navas-Castillo, J., Lau, D., Zerbini, F.M., 2015. Host range of *Euphorbia yellow mosaic virus* and its associated alphasatellite. Virus Rev Res 20 (Suppl), 198-198.

Nogueira, A.M., Nascimento, M.B., Barbosa, T.M.C., Quadros, A.F.F., Gomes, J.P.A., Orilio, A.F., Barros, D.R., Zerbini, F.M., 2021. The association between New World alphasatellites and bipartite begomoviruses: Effects on infection and vector transmission. Pathogens 10, 1244.

Pappas, N., Roux, S., Hölzer, M., Lamkiewicz, K., Mock, F., Marz, M., Dutilh, B.E., 2021. Virus Bioinformatics, in: Bamford, D.H., Zuckerman, M. (Eds.), Encyclopedia of Virology, 4th ed, pp. 124-132.

Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B.L., Salazar, Gustavo A., Bileschi, Maxwell L., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, Daniel H., Letunić, I., Marchler-Bauer, A., Mi, H., Natale, Darren A., Orengo, Christine A., Pandurangan, Arun P., Rivoire, C., Sigrist, C.J.A., Sillitoe, I., Thanki, N., Thomas, P.D., Tosatto, S.C.E., Wu, Cathy H., Bateman, A., 2023. InterPro in 2022. Nuc Acids Res 51, D418-D427.

Peeters, M.K.R., Menschaert, G., 2020. The hunt for sORFs: A multidisciplinary strategy. Exp Cell Res 391, 111923.

Piovesan, D., Necci, M., Escobedo, N., Monzon, A.M., Hatos, A., Mičetić, I., Quaglia, F., Paladin, L., Ramasamy, P., Dosztányi, Z., Vranken, W.F., Davey, Norman E., Parisi, G., Fuxreiter, M., Tosatto, Silvio C.E., 2020. MobiDB: intrinsically disordered proteins in 2021. Nuc Acids Res 49, D361-D367.

Rojas, M.R., Macedo, M.A., Maliano, M.R., Soto-Aguilar, M., Souza, J.O., Briddon, R.W., Kenyon, L.A., Rivera-Bustamante, R.F., Zerbini, F.M., Adkins, S., Legg, J.P., Kvarnheden, A., Wintermantel, W.M., Sudarshana, M.R., Peterschmitt, M., Lapidot, M., Martin, D.P., Moriones, E., Inoue-Nagata, A.K., Gilbertson, R.L., 2018. World management of geminiviruses. Annu Rev Phytopathol 56, 637-677.

Ruiz-Orera, J., Albà, M.M., 2019. Translation of small Open Reading Frames: Roles in regulation and evolutionary innovation. Trends Genet 35, 186-198.

Saunders, K., Bedford, I.D., Briddon, R.W., Markham, P.G., Wong, S.M., Stanley, J., 2000. A unique virus complex causes *Ageratum* yellow vein disease. Proc Natl Acad Sci USA 97, 6890-6895.

Saunders, K., Stanley, J., 1999. A nanovirus-like DNA component associated with yellow vein disease of *Ageratum conyzoides*: Evidence for interfamilial recombination between plant DNA viruses. Virology 264, 142-152.

Schlesinger, D., Elsässer, S.J., 2022. Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins. FEBS J 289, 53-74.

Tamura, K., Stecher, G., Kumar, S., 2021. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol 38, 3022-3027.

Xue, B., Williams, R.W., Oldfield, C.J., Goh, G.K., Dunker, A.K., Uversky, V.N., 2010. Viral disorder or disordered viruses: do viral proteins possess unique features? Protein Pept Lett 17, 932-951.

Zhou, X., 2013. Advances in understanding begomovirus satellites. Annu Rev Phytopathol 51, 357-381.