JOSÉ CLEYDSON FERREIRA DA SILVA

DATA WAREHOUSE ENRIQUECIDO COM MÉTODOS DE APRENDIZADO DE MÁQUINA PARA A FAMÍLIA GEMINIVIRIDAE

Dissertação apresentada a Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS - BRASIL
2016

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
S586d
2016

Silva, José Cleydson Ferreira da, 1985-
Data warehouse enriquecido com métodos de aprendizado de máquina para a família Geminiviridae : . / José Cleydson Ferreira da Silva. – Viçosa, MG, 2016.
xi,76f : il. (algumas color.) ; 29 cm.

Inclui apêndices.
Orientador: Fábio Ribeiro Cerqueira.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Aprendizado de máquina. 2. Bioinformática. 3. Biologia molecular. 4. Geminivírus. I. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-graduação em Ciência da Computação. II. Título.

CDD 22 ed. 006.31

JOSÉ CLEYDSON FERREIRA DA SILVA

DATA WAREHOUSE ENRIQUECIDO COM MÉTODOS DE APRENDIZADO DE MÁQUINA PARA A FAMÍLIA GEMINIVIRIDAE

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae.*

APROVADA: 25 de julho de 2016.

Sabrina de Azevedo Silveira

Francisco Murilo Zerbini Junior

Elizabeth Pacheco Batista Fontes
Coorientador

Fabio Ribeiro Cerqueira
Orientador

Dedico em especial à minha esposa Vivian e minha filha Luísa pelo amor, carinho e a cumplicidade que nos permitiram vencer juntos esta etapa;
Dedico aos meus pais João e Maria em agradecimento pelo dom da vida;
Aos meus irmãos Alexander (in memoriam), Sirlane, Sirlene, Cleydiane e Alexandre pelo convivio e amor fraterno.

Agradecimentos

À Deus agradeço pela saúde, família e amigos.
De maneira muito especial à minha esposa Vivian e minha filha Luísa, pelo incentivo, compreensão e encorajamento, durante todo este período;
Aos meus pais, por me educarem e me ensinarem com muita sabedoria valores que levarei por toda minha vida;
À toda família da minha esposa Vivian, pela amizade e acolhimento;
Aos meus irmãos Sirlane, Sirlene, Cleydiane e Alexandre, pela confiança e amizade;
À Universidade Federal de Viçosa, pela oportunidade de realização deste curso;
Ao meu orientador Prof. Fábio, pela disponibilidade, paciência, dedicação e profissionalismo;
À Prof. Elizabeth Fontes pela co-orientação, incentivo, oportunidades, dedicação e profissionalismo;
Aos amigos do Laboratório de Biologia Molecular de Plantas, Otavio, Pedro, João paulo, Maiana, Paola, Iara, Luiz, Marco Aurélio, Larissa, Laura, Bianca, Bruno, Dalton, Hanna, Otto, Igor,Virgílio, Gustavo Rosado, Michirito, Rafael, Thais, Valdeir, Amanda, Nayara e Maloni, pelo excelente convívio e troca de saberes;
Aos Professores Murilo Zerbini, Sérgio Bromenschenkel e Maximiller Dal-Bianco por auxiliarem nesta etapa e me acolherem em seus grupos de trabalho;
Aos amigos do Laboratório de Virologia Vegetal Murilo Alves, Márcio, Camila, Cesar, Talita, Angélica, Osvaldo e Larissa pelo excelente convívio e troca de saberes;
Em especial ao meu grande amigo Thales Francisco Mota Carvalho (Samwise Gamgee), pela amizade, competência, comprometimento e pelo ser humano

incrível que é.
Aos professores do DPI pela paciência, atenção e amizade;
Ao Prof. Fabiano Silva, pelo apoio para a realização deste trabalho e a amizade durante este tempo;
Ao Altino Alves pelo apoio e à amizade durante este tempo;
A todos que ajudaram e contribuíram direta ou indiretamente para a realização deste trabalho;
A gratidão é um dos únicos tesouros que podem tornar seres humanos mais humildes (Autor desconhecido).

Sumário

Lista de Tabelas **vii**

Resumo **viii**

Abstract **x**

1 Introdução **1**

2 Begomovirus Data Warehouse (BegomoDW): An integrated database for begomoviruses and related satellites **7**

3 Geminivirus Data Warehouse: A database enriched with machine learning approaches **18**

4 Fangorn Forest (F2): Um método para classificação de genes e gênero na família *Geminiviridae* **35**

5 Extração de informação de vírus de planta e importantes patógenos virais de humanos usando abordagens de processamento de linguagem natural **51**

6 Conclusões **66**

Referências Bibliográficas **68**

Apêndice A Arquivos suplementares do artigo 3 **72**

A.1 Arquivo suplementar 1 72

A.2 Arquivo suplementar 2 73

Apêndice B Trabalhos publicados ou em análise para publicação 75

Lista de Tabelas

A.1 Valor do IG dos atributos utilizados para obter os modelos de classificação de gênero. 72
A.2 Valor do IG dos atributos utilizados para obter os modelos de classificação de genes 73

Resumo

SILVA, José Cleydson Ferreira da, M.Sc., Universidade Federal de Viçosa, julho de 2016. **Data warehouse enriquecido com métodos de aprendizado de máquina para a família *Geminiviridae***. Orientador: Fabio Ribeiro Cerqueira. Coorientador: Elizabeth Pacheco Batista Fontes.

Geminivírus infectam uma ampla faixa de plantas monocotiledôneas e dicotiledôneas e causam expressivas perdas econômicas. A família *Geminividae* é uma das mais importantes famílias de vírus de plantas. Atualmente está composta por sete gêneros, é reconhecida pelo tipo de inseto vetor, hospedeiro, organização genômica e reconstrução filogenética. A amplificação por ciclo rolante permitiu que milhares de sequências completas e parciais fossem depositadas em bases de dados públicas. Entretanto, tais bases de dados são limitadas em ferramentas avançadas que permitam responder perguntas sofisticadas. Ao contrário de outros importantes patógenos virais, nenhum banco de dados para geminivírus que integre todas as informações relevantes foi ainda sugerido. Neste trabalho, um Data Warehouse (DW) designado geminivirus.com é proposto. Um DW amplamente enriquecido por abordagens de aprendizado de máquina que vise garantir confiabilidade e qualidade das sequências genômicas e seus metadados associados. As metodologias de extração, transformação dessas sequências e seus metadados foram implementadas em um processo ETL (Extract, Transform and Load) específico para dados de geminivírus. Além disso, neste processo, o uso de algoritmos de aprendizado de máquina como Multilayer Perceptron (MLP), Máquina de Vetores de Suporte (SVM) e Random Forest são utilizados como classificadores taxonômicos *in silico*

para classificar as sequências completas. Ademais, modelos de aprendizado de máquina foram propostos para classificação de genes. Os modelos para ambos os fins superam 98% de acurácia e precisão, utilizando apenas atributos extraídos da sequência genômica completa, sequência CDS (Coding DNA Sequence) e sequência de aminoácidos. Também técnicas de Processamento de Linguagem Natural baseadas em teoria dos grafos foram propostas para extração de informação e conhecimento em resumos de artigos. Essa metodologia apresentou grande potencial para responder perguntas específicas. Explorando o grafo de texto buscando por palavras chaves que representam os mecanismos evolutivos, verificou-se que o tema recombinação é os mais estudado se comparado à mutação, migração, seleção natural e deriva genética. Tornando-se assim, uma técnica propicia para gerar novas hipóteses. Ao utilizar tal técnica, observou-se que ferramentas de predição de genes não foram mencionadas. Dessa oportunidade, sugerimos um método para predição e classificação de genes designado Fangorn Forest (F2). Além disso, como parte desse método sugerimos um algoritmo para predição de genes designado Millau Bridge (MB). Esse algoritmo testa todas as possíveis ORFs que uma sequência genômica completa pode codificar por meio de codons de iniciação e terminação. Além disso, identifica sítios de excisão de splicing. geminivrus.com tornou-se uma base de dados robusta capaz de proporcionar dados com boa qualidade, ferramentas avançadas enriquecidas por métodos de aprendizado de máquina que auxiliam pesquisadores em suas atividades de pesquisa e tomada de decisão.

Abstract

SILVA, José Cleydson Ferreira da, M.Sc., Universidade Federal de Viçosa, July, 2016. **Data warehouse enriched with machine learning methods for the *Geminiviridae* family**. Adviser: Fabio Ribeiro Cerqueira. Co-adviser: Elizabeth Pacheco Batista Fontes.

Geminiviruses infect a wide range of monocot and dicot plants and cause significant economic losses. The Geminividae family is one of the most important plant virus families. Currently, it consists of seven genera and is recognized by the type of insect vector, host range, genome organization and phylogenetic reconstruction. The rolling cycle amplification allowed thousands of complete and partial sequences to be made available in public databases. However, such databases have limitations concerning advanced tools to answer sophisticated questions. Unlike other major viral pathogens, no database for geminiviruses that integrates all relevant information was suggested yet. In this work, a Data Warehouse (DW) designated geminivirus.org is proposed. It is a DW widely enriched by machine learning (ML) approaches designed to ensure reliability and quality of the genomic sequences and their associated metadata. The methods for extraction and transformation of these sequences and their metadata have been implemented using the ETL process (Extract, Transform and Load), specifically for geminivirus data. In addition, ML algorithms such as Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Random Forest classifier are used as in silico taxonomic classifiers to classify complete sequences. Furthermore, ML models are proposed for gene classification. All models exceed 98% accuracy and precision using only ex-

tracted attributes of the complete genome sequence, Coding DNA Sequence (CDS) and protein sequence. Additionally, Natural Language Processing based on graph theory techniques have been proposed for extracting information and knowledge articles. This methodology presented great potential to answer specific questions. While exploring the word graph by searching for keywords that represent evolutionary mechanisms, it was found that the subject of recombination is the most studied compared to the mutation, migration, natural selection and, genetic drift. The resulting method is demonstrated, thus, to be an interesting technique to generate new hypotheses. By using this technique, it was observed that gene prediction tools have not been mentioned. In this opportunity, we suggest a powerful method for prediction and classification of genes called Fangorn Forest (F2). Also as part of this method, we suggest a greedy algorithm for predicting genes designated Millau bridge (MB). This algorithm tests all possible ORFs that a complete genomic sequence can encode inspecting initiation and termination codons. Furthermore, it identifies splicing sites. geminivirus.org became a robust database capable of providing data with good quality, advanced tools enriched by machine learning methods that help researchers in their research activities and decision making.

Capítulo 1

Introdução

Plantas e vírus de plantas co-evoluem ao longo de milhões de anos. Geminivírus, nomeados assim por suas partículas icosaédricas geminadas, infectam diferentes tipos de plantas alimentícias, produtoras de fibras, ornamentais e plantas daninhas. O resultado dessa competição converte-se em severos impactos econômicos e danos expressivos à segurança alimentar no mundo todo. A família *Geminiviridae* é composta por sete gêneros: *Begomovirus*, *Mastrevirus*, *Curtovirus*, *Becurtovirus*, *Turncuntovirus*, *Eragrovirus* e *Topocovirus*. Além disso, alguns representantes do gênero *Begomovirus* estão associados com DNA satélites denominados alpha e betassatélites.

A descoberta da diversidade e dispersão geográfica de geminivírus tem sido possível devido ao esforço coletivo de muitos pesquisadores no mundo todo (Brown et al., 2015)(Basso et al., 2015), incluindo pesquisadores que visam encontrar meios de proporcionar imunidade às plantas (Zorzatto et al., 2015). O aprimoramento das técnicas de amplificação do genoma viral por ciclo rolante por meio da enzima phi29(Inoue-Nagata et al., 2004), bem como as técnicas de sequenciamento de DNA em larga escala e de baixo custo permitiram que milhares de sequências de DNA da família *Geminiviridae* fossem obtidas e disponibilizadas em bases de dados públicas. Além do sequenciamento desse material genético, também disponibilizou-se um imenso acervo de artigos científicos. Assim, para gerênciar o acervo de dados surgiram necessidades da utilização de grandes bases de dados específicas como Data Warehouse para reunir, relacionar

metadados e transformá-los em informação ou conhecimento. Além disso, surge a necessidade de agregar avançados algoritmos de aprendizado de máquina que sejam capazes de realizar a mineração de todos esses dados, tornando a geração de conhecimento um processo ágil e eficiente.

A utilização de Data Warehouse (DW) para gerenciar dados biológicos tem se tornado um desafio, uma vez que esses dados possuem natureza e fontes diversas e, em sua maioria, encontram-se disponíveis em bases de dados que não permitem responder perguntas específicas. Além disso, tais bases não foram projetadas para receber os dados de maneira padronizada e completa. Esse fato pode afetar a qualidade dos dados e permite que erros sejam disponibilizados e propagados.

Para iniciar as pesquisas do presente trabalho, optou-se em sugerir uma DW para o gênero *Begomovirus*, como descrito Capítulo 2. Begomovírus é o maior gênero entre os sete gêneros da família *Geminiviridae*. São conhecidos pelo elevado número de espécies e serem mono ou bipartidos. Os monopartidos possuem apenas o componente DNA-A, enquanto que os bipartidos apresentam os componentes DNA-A e DNA-B. Atualmente, estão disponíveis no banco de dados NCBI aproximadamente 13.000 sequências de DNA de begomovírus, sendo que muitas delas são sequências parciais, embora 6495 sequências são genomas completos.

Sugeriu-se algoritmos para construir ETL (Extract, Transform, Load ETL), para realizar a extração, transformação e carregamento de dados em uma DW designada BegomoDW. Os algoritmos incluem padronizar nomes de genes, sequências genômicas completas e definir de forma pontual ou aproximada a origem geográfica desses vírus. Além do mais, visam integrar todas as informações possíveis. Para isso, a modelagem de uma base de dados relacional foi projetada em diferentes dimensões. As tabelas SQL em modelo estrela (Boehnlein & Ulbrich-vom Ende, 1999) foram projetadas para relacionar o organismo como o centro da informação (tabela SQL central) e associar os metadados nas dimensões como sendo localização geográfica (lugar), data de coleta (tempo) e hospedeiro. Além disso, informações cruciais como material genético (DNA/proteína) e artigos científicos (*abstract*) foram relacionados para cada isolado (amostra coletada). Os mecanismos de busca implementados permitem a rápida recuperação de sequências e análises com as principais ferramentas de bioinformática, bem como demarcação de sequências baseados em comparações par a par e análises filogenéticas. Por fim,

disponibilizou-se um mapa interativo, que mostra cada ponto de coleta de amostras sequenciadas no mundo. Portanto, o BegomoDW de maneira geral tornou-se uma base de dados com elevado potencial de ampla utilização e expansão para agregar outros gêneros da família *Geminiviridae*. Os algoritmos implementados foram desenhados baseados em regras e em sua maioria conseguiram atuar nas regras estabelecidas com sucesso, embora não obtevessem sucesso quando as sequências de DNA e seus metadados associados apresentavam erros conceituais de submissão no Genbank (www.ncbi.nlm.nih.gov/). Identificou-se um elevado número de inconsistências nos dados inseridos em tal base de dados. Percebeu-se que a causa dessa limitação foi o fato de a aplicação de algoritmos baseados em regras para padronização e normalização serem limitados em detectar e corrigir erros de anotação causados pelo manuseio humano. As correções desses erros necessitam de abordagens mais robustas como aprendizado de máquina para detectar padrões por meio de aprendizado de máquina.

Para suprir essas necessidades e, ainda, tornar a DW mais completa, primeiramente ampliou-se o DW com a finalidade de receber dados dos sete gêneros da família *Geminiviridae* e seus DNAs satélites relacionados, como pode ser visto no Capítulo 3. Além disto, associaram-se métodos de aprendizado de máquina para auxiliar na padronização dos dados no momento em que as sequências são introduzidas no DW. Inicialmente, modelos robustos de aprendizado de máquina atuam como classificador taxonômico *in silico* por meio de características extraídas apenas das sequências virais. Em outras etapas, os genes são classificados utilizando atributos das sequências do DNA e sequências de proteínas. A fim de determinar modelos robustos, implementações de algoritmos de redes neurais (Gardner & Dorling, 1998), máquina de vetores de suporte (Platt, 1999) e árvores de decisão (Breiman, 2001) foram utilizadas. Essas técnicas têm sido aplicadas em dados biológicos para diversos fins (Frank et al., 2004) (Wang et al., 2005). Avaliando-se os modelos de classificação de predição de gênero e genes, ambos os casos superam 98% de acurácia, precisão e sensibilidade. Essa solução proposta, além de solucionar as restrições anteriormente mencionadas, auxiliou na confiabilidade e qualidade dos dados. Além disso, melhorias foram agregadas para gerenciar o fluxo e o histórico de informações no DW. Informações adicionais a respeito das sequências também foram acrescentadas. Informações sobre data de

coleta, hospedeiro e localização geográfica não foram disponibilizadas no momento em que as sequências foram submetidas. Tais informações foram acrescentadas por curadores que exploraram o imenso acervo científico e recuperaram as mesmas. Neste processo, resumos de artigos, livros (como o livro de taxonomia do ICTV - www.ictvonline.org/), artigos científicos completos e material suplementar foram manualmente consultados. Como resultado, logrou-se um elevado número de informações acrescentadas e disponibilizadas livremente.

Neste processo, observou-se que o acervo de resumos e artigos possuíam grande potencial de extração de informação se aplicadas técnicas de mineração. Com isto, as ferramentas computacionais resultantes serviriam para assessorar o processo de curadoria e ganho de informação no DW. Além dessas vantagens, verificou-se o potencial de extrair e relacionar informações dentro de estudos de geminivírus e compará-los com outros importantes patógenos virais, como aqueles que causam danos à saúde humana, tais como, ebola e o Zika vírus.

O processamento de texto para extração de informação é uma tarefa crítica. Na maioria das vezes, é baseada em contagem de palavras e aplicação de algumas métricas estatísticas(Martin & Jurafsky, 2000). Com estes recursos, há uma grande dificuldade para que o sentido semântico e o contexto que são empregados no texto sejam mantidos. As aplicações de Processamento de Linguagem Natural (NLP) baseado em teoria dos grafos têm sido empregadas para solucionar essas restrições. Além disso, NLPs propiciam análises textuais avançadas como análise de sentimento, reconhecimento de fala, reconhecimento de entidades e extração de informação (Bird et al., 2009)(Daniel & James, 2009). Tal fato, impulsionou a aplicação de NLP baseando-se em teoria dos grafos (Mihalcea & Radev, 2011) para extrair informações no acervo de geminivírus, trabalho este descrito no Capítulo 5. Resumidamente, construiu-se um grafo de texto, a partir de textos científicos, onde cada palavra de uma frase tornou-se um nó do grafo e a ligação de um nó ao outro se deu pela palavra subsequente, preservando a ordem original do texto. Assim, construiu-se um imenso grafo de palavras e utilizaram-se métricas de centralidade para avaliar a importância com que essas palavras foram empregadas. Utilizaram-se algoritmos para computar o número de conexões e a importância que cada palavra possui ao conectar a outras palavras na rede do grafo de texto. O algoritmo Google PageRank é o principal algoritmo do motor

de buscas do google.com. Sua utilização em nossas pesquisas foi empregada para identificar as palavras mais influentes nos textos escritos pelos pesquisadores com o objetivo de ranquear os contextos de pesquisas mais influentes e a maneira com eles se relacionam na rede. Além desse, o algoritmo betweenness centrality foi aplicado para identificar a centralidade do nó em meio a todos os caminhos mais curtos do grafo. A centralidade de graus também foi mensurada.

A análise da rede permitiu realizar inferências a respeito de palavras quando observadas por meio de contexto. Utilizaram-se como estudo de caso cinco mecanismos evolutivos para identificar ou corroborar com a literatura quais deles são os mais estudados. Além disso, visou-se inferir quais deles mais contribuem para a alta diversidade de espécies e a estrutura de suas populações. Os nós do grafo representados pelas palavras *mutation*, *recombination*, *migration*, *drift*, *natural* e *selection* foram visitados e o valor de suas métricas de centralidade foram examinados. O consenso dessa avaliação nos apontou que os eventos de recombinação são os mais frequentemente estudados, sugerindo que estes atuam de maneira significativa em relação aos demais.

Em todo o processo de desenho e construção dos algoritmos durante o desenvolvimento e testes das bases de dados BegomoDW e GeminvirusDW, observaram-se diversos erros em anotação de sequências de DNA e dos genes. Então, surgiu a necessidade de identificar quais ferramentas de bioinformática pesquisadores utilizam mais frequentemente para realizar o processo de curadoria desses dados. Para isso, utilizou-se o potencial do processamento de linguagem natural descrito acima, visitando-se o grafo de texto tendo como alvo as palavras machine, learning, blast, sdt, phylogenetic, rdp, rdp4, orf, finder. Com isto, observou-se que técnicas de machine learning ainda não haviam sido aplicadas em análise de dados de sequências da família *Geminiviridae*. Além disso, a principal ferramenta de predição de genes, ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html), não foi mencionada. Daí, notou-se que tais erros não seriam solucionados apenas com definições de regras triviais ou ferramentas existentes. Assim, sugeriu-se uma metodologia e uma ferramenta simples, rápida e amigável baseada em aprendizado de máquina para predizer genes e classificar ORFs e também detectar inconsistências nas sequências disponibilizadas publicamente.

O resultado foi a crição de um método que prediz e classifica genes designado Fangorn Forest (F2), descrito no Capítulo 4. O F2 realiza um pipeline completo desde a identificação do gênero e demarcação de espécies até a realização da construção filogenética de sequências virais de geminivírus. Para tarefa de classificação, 8 modelos foram gerados, sendo 1 para classificação do gênero e de satélites e outros 7 modelos específicos para classificação de genes, onde cada modelo é específico para cada gênero. Em testes de validação cruzada em 10 partes os modelos obtiveram 98% de precisão. É apresentado, ainda, o algoritmo designado Millau Bridge (MB) para predição de genes como parte do método F2, visando encontrar sítios de splicing em genes codificados por geminivírus (Wright et al., 1997). Tais métodos mostraram-se flexíveis para a aplicação em vírus com genoma circular menos complexos.

A utilização de DW, técnicas de integração de dados de diferentes fontes (ETL) e abordagens de aprendizado de máquina permitiram adquirir qualidade e confiabilidade nos dados aqui mencionados. A qualidade dessas sequências e informações possibilitam que complexos estudos em evolução como filogeografia, estrutura de populações, mutação e recombinação e taxonomia possam ser realizados com dados de boa qualidade. Além disso, proporcionam acesso rápido às informações de qualidade com usabilidade flexível e amigável. A interface amigável do DW, foi proporcionada com uso de tecnologias robustas como linguagem de programação Java, bases de dados MYSQL e tecnologias web baseadas em Java Script. Tais tecnologias facilitam a apresentação e interpretação gráfica de resultados por meio de gráficos, mapas e visualização da organização genômica de geminivírus, sugeridos por Varsani (Varsani et al., 2014). As abordagens de ML, aqui apresentadas, mostraram-se extremamentes eficiente para solucionar descuidos ocorridos no manuseio humano dos dados das espécies vírais alvos do presente estudo.

Neste trabalho, cada capítulo apresenta a evolução das ideias que aqui foram expostas e é constituído de um artigo científico que apresenta as metodologias propostas para solucionar cada dificuldade já mencionada.

Capítulo 2

Begomovirus Data Warehouse (BegomoDW): An integrated database for begomoviruses and related satellites

Artigo científico submetido ao CLEI 2016.

DATABASE

Begomovirus Data Warehouse (BegomoDW): An integrated database for begomoviruses and related satellites

Jose Cleydson F Silva[1,2], Marcos F Basso[4,2], Thales F M Carvalho[1], Michihito Deguchi[2], Welison A Pereira[2], Roberto R Sobrinho[4,2], Otávio JB Brustolini[3,2], Pedro M P Vidigal[2], Alison T M Lima Lima[4,2], Anésia A Santos Santos[3,2], Maximiller D L Costa[3], Francisco Murilo Zerbini[4], Fabio R Cerqueira[1] and Elizabeth B P Fontes[3,2*]

*Correspondence: bbfontes@ufv.br
[1]Department of Informatics, Universidade Federal de Vicosa, Campus universitario, Vicosa, BR
Full list of author information is available at the end of the article

Abstract

Begomoviruses (whitefly-transmitted geminiviruses) are plant single-stranded DNA viruses that are responsible for serious economic losses worldwide and are frequently associated with satellite DNAs (alphasatellites and betasatellites). We developed a database that integrates all relevant information and provides user-friendly tools to easily manipulate begomovirus-related data. Here, we describe a novel data warehouse comprising integrated information of begomoviruses and associated DNAs. The data warehouse was implemented using an MVC software architecture pattern written in the Perl programming language and MariaDB SQL server. The data were based on data retrieved from the GenBank and PubMed web repositories. We developed Perl ETL Toolkit for extract data and web interface which contains search modules for viral species, full-length genomes, protein and gene sequences as well as scientific publications. We incorporated useful tools, such as a BLAST, identity matrix comparisons between sequences, phylogenetic analysis and the visualization of related information based on geographical distribution. The BegomoDW provides a user-friendly environment to retrieve information about begomoviruses and related satellite DNAs. The BegomoDW is freely available at inctipp.bioagro.ufv.br/begomodw.

Keywords: Data Warehouse; Begomoviruses; ETL

Introduction

Begomoviruses are plant single-stranded DNA viruses that infect dicotyledonous plants and are responsible for expressive economic losses worldwide [1]; [2]; [3]; [4]; [5]. They belong to the *Geminiviridae* family and are whitefly [Bemisia tabaci Middle East-Asia Minor 1 species (MEAM1), formerly known as the B biotype]-transmitted viruses. The diseases caused by begomoviruses have economically and socially impacted most continents: Europe and Asia (tomato and cotton leaf curl diseases, TYLCD and CLCuD, respectively), Africa (the cassava mosaic disease, CMD) and the Americas (bean golden mosaic disease, BGMD; [6]). Begomoviruses can be monopartite (single genomic DNA of approximately 2.9 kb) and bipartite (two DNA components, each approximately 2.6 kb, referred to as DNA-A and DNA-B; [7]). Both virion-sense and complementary-sense strands are transcribed and encode several viral proteins ([8]). Typically, the "Old World Begomoviruses" (Europe, Asia

and Africa) are primarily monopartite and associated with satellite DNAs (alpha and betasatellites [9]), whereas the "New World Begomoviruses" (Americas) are primarily bipartite and may be associated with satellite[10][11]. The DNA-A of bipartite begomoviruses contains genes involved in the replication and encapsulation of the viral genome, as well as the RNA silencing suppression, whereas the DNA-B encodes proteins associated with intra- and intercellular viral movement [12]. The genomic components do not share significant identity, with the exception of approximately 200 bp of a common region, which harbors the geminivirus conserved nanonucleotide (TAATATTAC) sequence and the replication origin [13][14].

Monopartite begomoviruses have a genomic organization similar to that of DNA-A [9]. Betasatellites are approximately 1.35 kb in size and contain a single Open Reading Frame (ORF) that encodes a pathogenicity determinant protein. This protein is primarily responsible for the development of typical symptoms, the enhancement of helper-begomovirus pathogenicity and the modulation of the virus host range by interfering with the host defense response [15][16]. Alphasatellites have a genomic size similar to that of the betasatellites and the DNA-R component of nanoviruses (1.0 kb), which contain only an ORF. The ORF encodes replication-associated protein (Rep) and does not significantly contribute to disease development. These satellites can autonomously replicate but require a helper virus for host systemic infection and transmission through an insect vector [17].

Recent research reports a high frequency of begomovirus-satellite virus complexes occurrence [10][11]. Currently, more than 288 begomovirus species and numerous species of begomovirus-associated alpha and betasatellites have been described [18][19]. The current high-throughput sequencing methods and metagenomics approaches have accelerated virome knowledge and facilitated the identification of several viral agents in a wide variety of samples [20][21]. Due to the advent of rolling-circle amplification (RCA) with phi-29 DNA polymerase [22], thousands of full-length begomovirus and satellite genomic sequences have become available in public databases. Prior to 2003, only 25 full-length genomic sequences of the begomovirus DNA-A or DNA-B were deposited or available in the GenBank database from the National Center for Biotechnology Information (NCBI). Despite there is not an official statistics about begomovirus sequences in GenBank database, it has been retrieved over 5,750 begomovirus full-length genomic sequences (DNA-A, DNA-B and monopartite genomes) and 854 satellite sequences. To deal with this large amount of data, it is necessary a time-consuming task and relative high level of computational expertise.

Databases that integrate all relevant information and provide user-friendly tools to easily manipulate begomoviruses-related data remain to be developed. To address this restriction, we herein describe a novel database designated the Begomovirus Data Warehouse (BegomoDW), which comprises integrated information on begomoviruses and their related satellite DNAs. This system is also enriched with functionalities that facilitate bioinformatics analysis. BegomoDW provides useful tools, such as the ability to search for any sequences (full-length genomes, genes or proteins) of begomoviruses and related satellite DNAs that are associated with precise or approximate geographical coordinates, a basic local alignment search, a pairwise comparison and identity matrix between sequences and an optimized phylogenetic analysis.

Methods

Data loading and integration

BegomoDW was implemented using an MVC software architecture pattern (Model-View-Controller) with modules programmed in the Perl programming language, MariaDB SQL server relational databases and an Apache Web Server. The data structure was organized with SQL tables based on data retrieved from GenBank (http://www.ncbi.nlm.nih.gov/genbank) and PubMed (http://www.pubmed.com) from NCBI web repositories. Our custom scripts parse the data related to begomoviruses and related satellite DNAs and insert them into the system using the Perl Toolkit ETL (Extract, Transform and Load). The design and workflow of BegomoDW are summarized in Figure 1.

Data source

BegomoDW data were retrieved from GenBank, PubMed and Google Maps. The viral sequences were filtered by definition, organism name (begomoviruses, alpha or betasatellites) and sequence length to enable customized searches. The scientific publications were also retrieved based on common descriptions from the PubMed NCBI database (Figure 1).

Data curation

A manual curation of the data was performed prior to data insertion and their updates into BegomoDW. In silico curation was also implemented using algorithms such as that for correcting genomic sequences that do not start with the 5' conserved cytosine of the Rep cleavage site. For old genomic sequences, the date of collection, geographic location and hosts are not available in GenBank. BegomoDW is open-access and enables the inclusion of information that may be provided by research groups responsible for these genomic sequences. The sequences from partial genomic and patent is not included in BegomoDW because they are not suitable for begomovirus taxonomy studies.

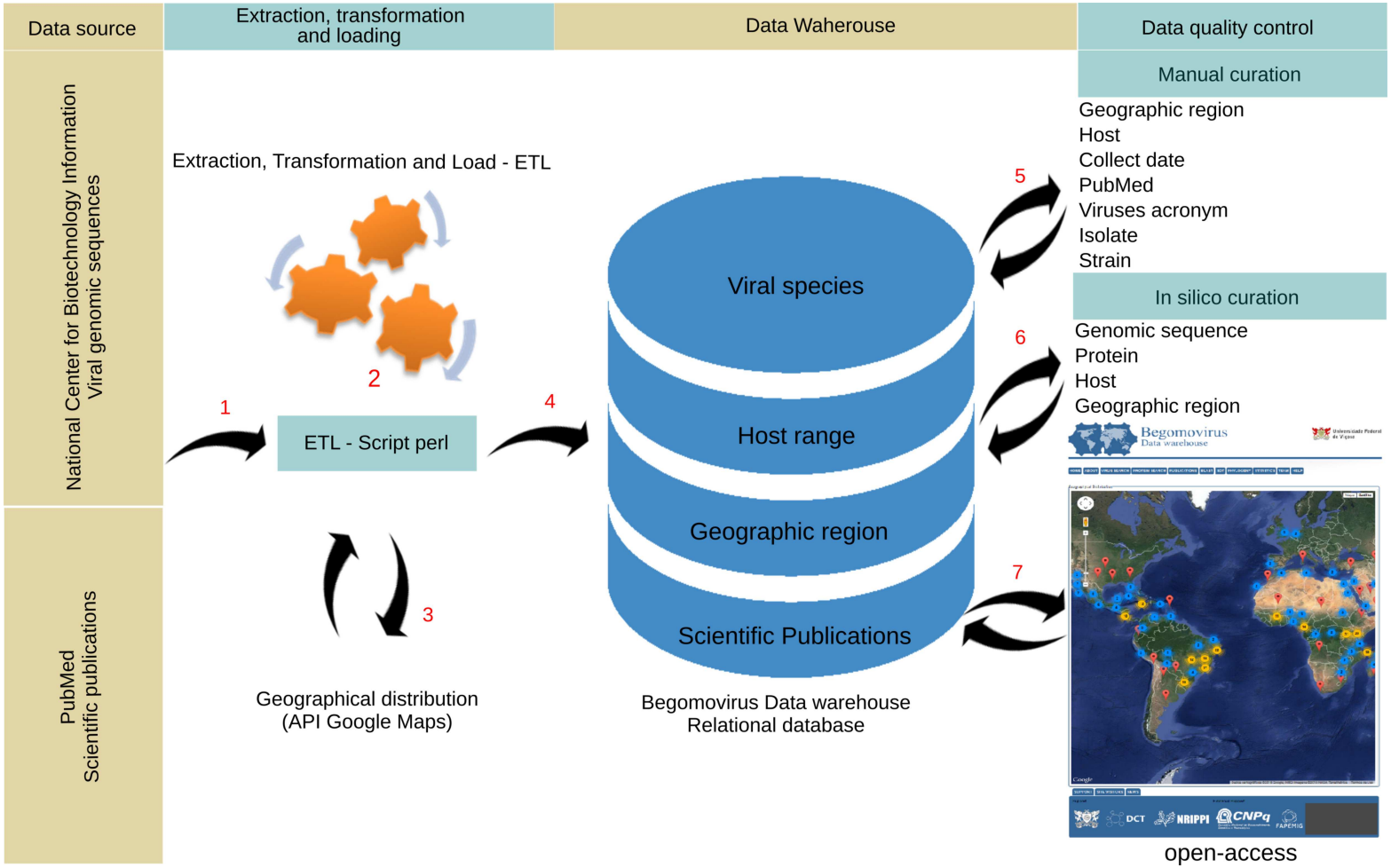

Figure 1 The design and workflow of BegomoDW. The data were retrieved from GenBank and PubMed (NCBI) web repositories based on common descriptors (1). The data were selected and transformed (2), and the geographic coordinates of the isolates were obtained using API Google Maps (3). The data were loaded and integrated in a relational database (4). They were then manually curated (5), and genomic sequences, the geographic region, protein description and host were corrected using several algorithms (6). Finally, the information and analytical tools were made available in open-access software without any restrictions to those outside the academic scientific community (7).

Analysis tools

The web interface contains optimized search modules for viral sequences and scientific publications. To perform a basic local alignment search with genomic, protein or CDS sequences, we embedded BLAST software (BLASTn, BLASTp and BLASTx algorithms) with default p-value parameters. The user can perform a search using key words such as viral name, host, accession number, country, genomic segment (DNA-A, DNA-B or satellite), collection year or submission year. The search for scientific publications is also performed using key words, such as PMID number, researcher name, virus name, journal name, publication year and others. BegomoDW also provides tools to perform pairwise comparisons and identity matrix between sequences, optimized phylogenetic analysis and the geographical visualization of begomovirus-related data.

We incorporated the software SDT v1.0 into BegomoDW [23]. This enables pairwise sequence comparison analyses. For these analyses, the query sequences are used to search for sequence identity in the BegomoDW database via MAFFT [24], MUSCLE [25] or CLUSTALW [26] algorithms. Based on the results of sequence length, coverage and identities, the user can select the desired sequences and generate a comparative matrix for identity. The identity matrix generated can be viewed in BegomoDW or downloaded. The matrix can be opened in SDT software and edited using any image editing software.

To automated phylogenetic analysis, the user initially enters at least one sequence (other begomovirus sequences can also be used) and then performs a search for sequence homology using the BLAST algorithm. The alignment output is automatically loaded into the FastTree2 software [27] for inference via maximum-likelihood phylogenies and default parameters with 1,000 bootstraps. The FastTree 2 software uses minimum-evolution subtree-pruning-regrafting and maximum-likelihood NNIs to search for better trees. For visualization and additional analysis of the phylogenetic trees in BegomoDW, we included the Phytools R package, which uses a fastBM simulation function [28]. The phylogenetic tree output can also be downloaded in Newick format and can be opened and edited, for example, using FigTree v1.4.2 open source software, which is available at http://tree.bio.ed.ac.uk/software/figtree.

All information related to begomoviruses and/or associated satellites, such as species and geographical distribution, can be visualized using a graphic interface developed with Google Maps API (https://developers.google.com/maps/?hl=en) and Google MarkerClusterer (https://github.com/googlemaps/js-marker-clusterer). Additional statistical information about the amount of the full-length genome sequences per country or per species, the amount of complete genome sequences per year and the amount of scientific publications related to begomoviruses and/or associated satellite DNAs per year are shown with charts using Google charts API (https://developers.google.com/chart/?hl=en).

Information updates

The information available in BegomoDW will be automatically updated weekly with GenBank and PubMed parsers using a program written in Perl and integrated in the BegomoDW code base. Other timely updates will be performed manually.

Information regarding viral genomic sequences unavailable on NCBI web parsers can be updated into BegomoDW as soon as the sequences are received from the respective research groups.

Utility and Discussion

Begomoviruses infect cultivated and non-cultivated plants worldwide [29]; [30]. The high molecular diversity, quantity of viral species and the large amount of sequences make difficult to recover, analize and interpret begomovirus data. The data search coud be inefficient, time-consuming and required high computational expertise. The number of full-length genomic sequences and scientific publications related to begomoviruses and their associated satellite DNAs has increased significantly over the last 10 years (Figure 2). This tendency is due at least in part to the technique of circular DNA amplification (through RCA) using DNA polymerase of bacteriophage phi29 [22] as well as novel DNA sequencing approaches [21]. A timeline developed by BegomoDW summarizes the extent of scientific inquiry on begomoviruses and related alpha and betasatellite genomic sequences that have been deposited in public databases (Figure 2). The timeline also highlights growth and major developments observed in recent years. However, unlike other important viral pathogens [Hepatitis C (`hcv.lanl.gov`) and HIV (`hiv.lanl.gov`)], which possess consolidated databases, no databases include all relevant begomovirus-related information. We implemented BegomoDW and integrated various levels and types of begomovirus-related information in an attempt to promote the rapid and easy recovery of data and information on descriptors from various web repositories (GenBank and PubMed).

BegomoDW is maintained by the National Institute of Science and Technology in Plant-Pest Interactions (INCTIPP) and the Laboratory of Plant Molecular Biology (LBMP) (`http://www.inctipp.ufv.br/en/`) located in the Institute of Biotechnology Applied to Agriculture (BIOAGRO) at the Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil.

Data source of begomoviruses and related satellite DNAs

BegomoDW enables users to query from their own web pages and applications. The available information was extracted from GenBank (viral sequences) and PubMed (scientific publications) databases and was parsed from the NCBI data warehouse (Figure 1).

Data quality control

The key point of BegomoDW is the information gain by manual inspection. A large number of genomic sequences available on GenBank are not associated with scientific publications (despite having been identified). Efforts to associate viral sequences with their respective scientific publications are often time-consuming and dependent on the curator's skills. Thus, relevant information, such as the host plant, collection site (including geographical coordinates when available), viral acronym, strain, isolate, date of sampling and associated scientific publications, was searched for each sequence in the database.

Begomovirus and related satellite virus genomic sequences

The search and analysis tools of BegomoDW provide various search criteria on both nucleotide sequences (entire genomes or CDS) and predicted protein sequences encoded by the viral genome. The data curation provides users with the ability to perform searches based on parameters (or combinations thereof) such as viral name, host plant, access number in GenBank, country or continent in which the sequences were sampled, or the genomic segment (DNA-A, DNA-B, monopartite begomoviruses, alpha and betasatellite DNAs).

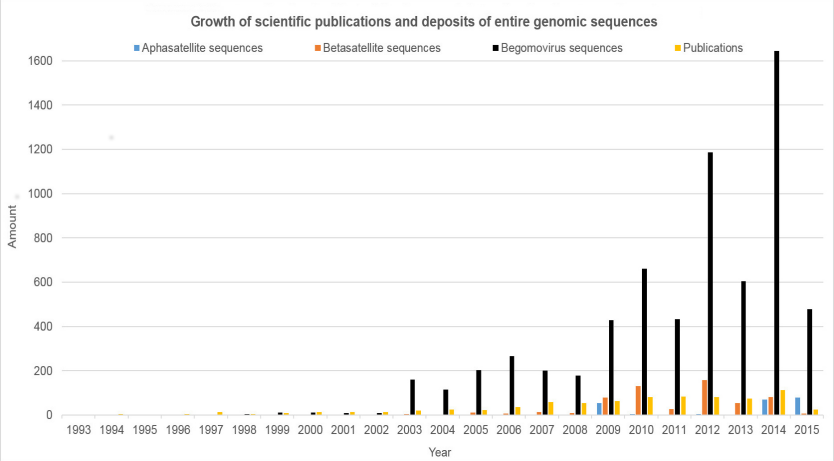

Figure 2 *

Growth of scientific publications and deposits of full-length genomic sequences. A timeline developed by the Begomovirus Data Warehouse (BegomoDW) summarizes the ammount of scientific publications and the deposits of full-length genomic sequences of begomoviruses and alpha and betasatellite viruses into a public database in recent years.

Scientific publications

All scientific publications related to sequences deposited into GenBank can be easily recovered from BegomoDW. Data curation on BegomoDW provides integration of the sequences deposited into GenBank with the associated scientific publication. Thus, BegomoDW provides the most complete information regarding begomovirus sequences.

Geographical distribution of begomoviruses and related satellite DNAs

Genomic sequences with geographical coordinates or other sampling locations (state or country) were used to provide information on the worldwide geographical distribution of begomoviruses and their related satellite DNAs. For the cases that lack precise geographic coordinates from GenBank, we recovered more generic information with respect to sampling localization and represented them as widespread coordinates. Therefore, the user can obtain a list of species associated with a given geographical location. This geographic data representation provides a convenient tool for rapid data recovery (e.g., for phylogeography studies).

BegomoDW features

Users can perform a basic local alignment search against a query sequence (nucleotide or amino acid) using BLAST software embedded in BegomoDW. The BLAST serves as a tool for initial sequence search with high similarity. These results can be automatically used in the subsequent analysis via SDT v1.0 software [23] implemented in BegomoDW. The SDT tool was recently implemented for viral species demarcation and provides a standardization for all parameters, such as alignments and for processing gaps, to calculate the identity values between full-length viral genomes or gene sequences. We incorporate an adapted and parallelized version of the SDT software into BegomoDW. This enables genomic sequences of begomoviruses and related satellite DNAs to be directly compared and eliminates the need for a pre-installed SDT version on the user's computer. The analysis briefly performs a preliminary comparison of the query sequence with other available sequences using BLAST software. This enables a pre-selection of the most related sequence. Then, the SDT software performs all comparisons between the sequence queries provided by the user and those that were pre-selected in the previous step of the BLAST results.

Another advantage of using the SDT software from BegomoDW is that the algorithm only performs comparisons involving query sequences provided by the user against those available in BegomoDW. This reduces the analysis complexity and duration needed to generate results. Thus, analyses involving a large number of sequences may be performed within a short duration (especially if compared with analyses performed on the user's computer). Only one SDT version for Windows is freely available for download. Its implementation into BegomoDW enables users of various platforms to use this software. Finally, the user can obtain a color array that represents the identity values that can be downloaded to the user's computer as a list containing the results of all of these pairwise comparisons.

The phylogenetic analysis from BegomoDW enables a rapid visualization of the phylogenetic relationships and groupings from the input sequence dataset. The sequence alignments are performed using the MUSCLE software, and the output is automatically loaded into the FastTree 2 software to optimize the analysis. FastTree 2 is a tool that enables the inference of phylogenies for alignments with up to hundreds of sequences. Thus, the sequence number used is not a limiting factor for this analysis. FastTree 2 is also slightly more accurate than its former version and 100-1,000-fold faster than other tools [27].

Conclusions

BegomoDW is an integrated and open-access data warehouse that optimizes complicated and comprehensive searches that are difficult to perform using existing tools. Therefore, BegomoDW efficiently assists in targeted information search and provides accurate and concise information on all begomoviruses and related satellite DNAs to the scientific community. Its flexibility enables the addition or analysis of various types of taxonomic, genomic, sampling or biological data to facilitate and update information sources. The data sources and additional analytical tools will greatly facilitate searches of the BegomoDW information management system. This open-access software will not be restricted in use to academics and will represent a valuable resource for a broad research community.

BegomoDW will be monthly updated with new information extracted from scientific publications, meetings and abstracts. The inclusion of new data sources will contribute to a better information currently available to accommodate further information that can assist in the interpretation of bioinformatics analysis results. We will update all pertinent information of begomoviruses and related satellite DNAs, such as new genomic sequences, taxonomy, novel criteria for demarcating species, scientific publications and recent advances. The users will also actively contribute by sending feedback, minor corrections and suggestions to begomodw@ufv.br. Future improvements will permit the further development of meta-analyses. The BegomoDW is freely available at inctipp.bioagro.ufv.br/begomodw.

Abbreviations
CDS: Coding DNA Sequence **ETL:** Extract ,Transform, Load **SDT:** Sequence Demarcation Tools

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
The authors are grateful to the National Institute of Science and Technology in Plant-Pest Interactions (INCT-IPP), the Fundação de Amparo à Pesquisa do estado de Minas Gerais (FAPEMIG), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for financial support.

Author details
[1]Department of Informatics, Universidade Federal de Vicosa, Campus universitario, Vicosa, BR. [2]Instituto Nacional de Ciência e Tecnologia em Interações Planta-Praga/BIOAGRO, Campus Universitário, Vicosa, Brazil. [3]Departamento de Bioquímica e Biologia Molecular, Campus Universitario, Vicosa, Brazil. [4]Departamento de Fitopatologia, Campus Universitario, Vicosa, Brazil.

References

1. Barbosa, L.d.F., Marubayashi, J.M., De Marchi, B.R., Yuki, V.A., Pavan, M.A., Moriones, E., Navas-Castillo, J., Krause-Sakate, R.: Indigenous american species of the Bemisia tabaci complex are still widespread in the americas. Pest management science **70**(10), 1440–1445 (2014)
2. Briddon, R.W.: Cotton leaf curl disease, a multicomponent begomovirus complex. Molecular Plant Pathology **4**(6), 427–434 (2003)
3. Monci, F., Sánchez-Campos, S., Navas-Castillo, J., Moriones, E.: A natural recombinant between the geminiviruses Tomato yellow leaf curl sardinia virus and Tomato yellow leaf curl virus exhibits a novel pathogenic phenotype and is becoming prevalent in spanish populations. Virology **303**(2), 317–326 (2002)
4. Morales, F.J., Anderson, P.K.: The emergence and dissemination of whitefly-transmitted geminiviruses in latin america. Archives of virology **146**(3), 415–441 (2001)
5. Were, H., Winter, S., Maiss, E.: Viruses infecting cassava in kenya. Plant disease **88**(1), 17–22 (2004)
6. Wyant, P.S., Strohmeier, S., Schäfer, B., Krenz, B., Assunção, I.P., de Andrade Lima, G.S., Jeske, H.: Circular DNA genomics (circomics) exemplified for geminiviruses in bean crops and weeds of northeastern brazil. Virology **427**(2), 151–157 (2012)
7. Briddon, R.W., Patil, B.L., Bagewadi, B., Nawaz-ul-Rehman, M.S., Fauquet, C.M.: Distinct evolutionary histories of the DNA-a and DNA-b components of bipartite begomoviruses. BMC evolutionary biology **10**(1), 97 (2010)
8. Hanley-Bowdoin, L., Bejarano, E.R., Robertson, D., Mansoor, S.: Geminiviruses: masters at redirecting and reprogramming plant processes. Nature Reviews Microbiology **11**(11), 777–788 (2013)
9. Mansoor, S., Briddon, R.W., Zafar, Y., Stanley, J.: Geminivirus disease complexes: an emerging threat. Trends in plant science **8**(3), 128–134 (2003)
10. Paprotka, T., Metzler, V., Jeske, H.: The first DNA 1-like α satellites in association with new world begomoviruses in natural infections. Virology **404**(2), 148–157 (2010)
11. Romay, G., Chirinos, D., Geraud-Pouey, F., Desbiez, C.: Association of an atypical alphasatellite with a bipartite new world begomovirus. Archives of virology **155**(11), 1843–1847 (2010)
12. Rojas, M.R., Hagen, C., Lucas, W.J., Gilbertson, R.L.: Exploiting chinks in the plant's armor: evolution and emergence of geminiviruses. Annu. Rev. Phytopathol. **43**, 361–394 (2005)
13. Fondong, V.N.: Geminivirus protein structure and function. Molecular plant pathology **14**(6), 635–649 (2013)
14. Hanley-Bowdoin, L., Settlage, S.B., Orozco, B.M., Nagar, S., Robertson, D.: Geminiviruses: models for plant DNA replication, transcription, and cell cycle regulation. Critical Reviews in Plant Sciences **18**(1), 71–106 (1999)
15. Saeed, M., Behjatnia, S.A., Mansoor, S., Zafar, Y., Hasnain, S., Rezaian, M.A.: A single complementary-sense transcript of a geminiviral DNA β satellite is determinant of pathogenicity. Molecular Plant-Microbe Interactions **18**(1), 7–14 (2005)
16. Saunders, K., Norman, A., Gucciardo, S., Stanley, J.: The DNA β satellite component associated with ageratum yellow vein disease encodes an essential pathogenicity protein (βc1). Virology **324**(1), 37–47 (2004)

17. Briddon, R., Stanley, J.: Subviral agents associated with plant single-stranded DNA viruses. Virology **344**(1), 198–210 (2006)
18. Fauquet, C., Briddon, R., Brown, J., Moriones, E., Stanley, J., Zerbini, M., Zhou, X.: Geminivirus strain demarcation and nomenclature. Archives of virology **153**(4), 783–821 (2008)
19. Kumar, J., Kumar, J., Singh, S.P., Tuli, R.: Association of satellites with a mastrevirus in natural infection: complexity of wheat dwarf india virus disease. Journal of virology **88**(12), 7093–7104 (2014)
20. Rwahnih, M.A., Dave, A., Anderson, M.M., Rowhani, A., Uyemoto, J.K., Sudarshana, M.R.: Association of a DNA virus with grapevines affected by red blotch disease in California. Phytopathology **103**(10), 1069–1076 (2013)
21. Radford, A.D., Chapman, D., Dixon, L., Chantrey, J., Darby, A.C., Hall, N.: Application of next-generation sequencing technologies in virology. Journal of General Virology **93**(9), 1853–1868 (2012)
22. Inoue-Nagata, A.K., Albuquerque, L.C., Rocha, W.B., Nagata, T.: A simple method for cloning the complete begomovirus genome using the bacteriophage φ29 DNA polymerase. Journal of virological methods **116**(2), 209–211 (2004)
23. Muhire, B., Martin, D.P., Brown, J.K., Navas-Castillo, J., Moriones, E., Zerbini, F.M., Rivera-Bustamante, R., Malathi, V., Briddon, R.W., Varsani, A.: A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus mastrevirus (family geminiviridae). Archives of Virology **158**(6), 1411–1424 (2013)
24. Katoh, K., Misawa, K., Kuma, K.-i., Miyata, T.: MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic acids research **30**(14), 3059–3066 (2002)
25. Edgar, R.C.: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research **32**(5), 1792–1797 (2004)
26. Thompson, J.D., Higgins, D.G., Gibson, T.J.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research **22**(22), 4673–4680 (1994)
27. Price, M.N., Dehal, P.S., Arkin, A.P.: FastTree 2–approximately maximum-likelihood trees for large alignments. PloS one **5**(3), 9490 (2010)
28. Revell, L.J.: phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution **3**(2), 217–223 (2012)
29. Brustolini, O.J., Machado, J.P.B., Condori-Apfata, J.A., Coco, D., Deguchi, M., Loriato, V.A., Pereira, W.A., Alfenas-Zerbini, P., Zerbini, F.M., Inoue-Nagata, A.K., *et al.*: Sustained NIK-mediated antiviral signalling confers broad-spectrum tolerance to begomoviruses in cultivated plants. Plant biotechnology journal **13**(9), 1300–1311 (2015)
30. Zorzatto, C., Machado, J.P.B., Lopes, K.V., Nascimento, K.J., Pereira, W.A., Brustolini, O.J., Reis, P.A., Calil, I.P., Deguchi, M., Sachetto-Martins, G., *et al.*: NIK1-mediated translation suppression functions as a plant antiviral immunity mechanism. Nature **520**(7549), 679–682 (2015)

Capítulo 3

Geminivirus Data Warehouse: A database enriched with machine learning approaches

Artigo científico a ser submetido à revista Bioinformatics Oxford.

DATABASE

Geminivirus Data Warehouse: A database enriched with machine learning approaches

Jose Cleydson F Silva[1,2], Thales F M Carvalho[1], Marcos F Basso[1], Michihito Deguchi[2], Welison A Pereira[2], Roberto R Sobrinho[2], Pedro M P Vidigal[3], Otávio JB Brustolini[2], Fabyano F Silva[5], Maximiller D L Costa[2], Anésia A Santos[2], Francisco Murilo Zerbini[6,2], Fabio R Cerqueira[1] and Elizabeth B P Fontes[2,4*]

*Correspondence: bbfontes@ufv.br
[1]Departamento de Informática, Universidade Federal de Viçosa, Campus Universitário, Viçosa, Brazil
Full list of author information is available at the end of the article
†Equal contributor

Abstract

Background: The *Geminiviridae* family encompasses a group of single-stranded DNA viruses with twinned and quasi-isometric virions that infect a wide range of dicotyledonous and monocotyledonous plants and are responsible for significant economic losses worldwide. Geminiviruses are divided into seven genera, according to insect vector, host range, genome organization, and phylogeny reconstruction. Using rolling-circle amplification approaches, thousands of full-length geminivirus and satellite genome sequences were amplified and along with the advancement of sequencing technologies have become available in public databases. As a consequence, many important challenges have emerged, namely, how to classify, store, and analyze massive data sets as well as how to extract information or new knowledge. Data mining approaches, mainly supported by machine learning techniques, are a natural means for high-throughput data analysis in the context of genomics, transcriptomics, proteomics, and metabolomics.

Results: Here, we implemented search modules, bioinformatics tools, and machine learning methods to retrieve high precision information, demarcate species, and create classifiers for genera and ORFs.

Conclusions: The development of a Data Warehouse enriched with machine learning approaches, designated geminirus.com, is described in the present investigation. The technical use of data mining, such as ETL (Extract, Transform, Load), algorithms based on machine learning and on rules allowed us to obtain a database with quality data and suitable tools for bioinformatics analysis. The Geminivirus Data Warehouse (geminivirus.org) offers a simple and user-friendly environment for information retrieval and knowledge discovery related to geminiviruses.

Keywords: Machine learning; Randon Forest; knowledge discovery; Data mining; Geminivirus

Background

The advancement of sequencing technologies enabled the rapid increasing of genomic data in public databases and introduced genomics in the era of massive data generation. The biggest challenges, thus, turned out to be how to acquire, classify, store and analyze huge data sets and extract knowledge from them [1]. Moreover, the process of massive data analysis has additional challenges, such as: how high volume is processed, how to speed up the process and how to maintain the data veracity.

To extract and process data of interest, it is recommended to use the Knowledge Discovery in Databases (KDD) process, by which the data are selected, preprocessed, transformed, mined, and evaluated [2][3]. The data mining step includes application of unsupervised and supervised methods such as cluster analysis, classification, and rule learning techniques [4]. Machine learning techniques and data mining applications have been suggested for high-throughput data analysis in plants as well for all levels of studies, i.e., in genomics, transcriptomics, proteomics, and metabolomics [5], including taxonomic classification in metagenomic data [6]. The current high-throughput sequencing methods, metagenomics analysis approaches, and powerful bioinformatics tools accelerated the knowledge of a number of viromes, allowing the identification of several viral agents in a wide range of cultivated and uncultivated plants. Furthermore, using rolling-circle amplification approaches, thousands of full-length geminivirus and satellites genome sequences have been amplified and have become available in public databases.

The *Geminiviridae* family is a group of single-stranded DNA (ssDNA) viruses, with twinned and quasi-isometric virions that infect a wide range of dicotyledonous and monocotyledonous plants, and are responsible for important economic losses in tropical and subtropical regions around the world. The *Geminiviridae* family is composed by seven genera: *Becurtovirus*, *Begomovirus*, *Curtovirus*, *Eragrovirus*, *Mastrevirus*, *Topocuvirus* and *Turncurtovirus*. The current classification is based on the insect vector, host range, genome organization, and phylogeny reconstruction [7] [8]. Viruses in the genus *Begomovirus* can be monopartite (single genomic DNA) or bipartite (two DNA components, referred to as DNA-A and DNA-B) and are exclusively transmitted by whitefly (*Bemisia tabaci*). The DNA-A contains genes required for DNA replication (*Rep*, *Ren*), gene expression control (*TrAP*), suppression of host defenses (*TrAP* and *AC4*), and viral genome encapsidation (*CP*), whereas the DNA-B encodes two proteins involved in intra- and intercellular movement (NSP and MP) [9]. The genus *Mastrevirus* is composed of viruses transmitted by leafhoppers, which infect a wide range of monocotyledonous plants and have a single genome component that encodes four proteins, a movement protein (pre-coat), a coat protein (*CP*), and two splicing variants of the replication-associated protein (*Rep*) [10]. Currently, the genus *Becurtovirus* has two viral species, *Beet curly top Iran virus*, transmitted by leafhopper, and *Spinach curly top Arizona virus* with unknown vector. The genomic structure of becurtoviruses contains four genes, a pre-coat gene, a *CP*, two Reps, and possibly a regulatory gene (*Reg*). Viruses in the genera *Eragrovirus* and *Turncurtovirus* are transmitted by unknown vectors. Viruses from both genera encode pre-coat protein, *CP*, *Rep*, and transactivator protein (*TrAP*). However, turncurtoviruses encodes two additional proteins, the replication enhancer (*Ren*) and symptom determinant proteins (*sd/p.sd*) [8]. The viruses belonging to the genus curtovirus are transmitted leafhoppers and their genomic structure is composed by seven genes: pre-coat gene, regulatory gene (*Reg*), *CP*, *Ren*, *TrAP*, *Rep*, and *sd/p.sd*. The curtoviruses genome structure is similar to begomoviruses, topocuviruses, and turncurtoviruses [11]. The genus *Topocuvirus* has only one genome sequence deposited in public databases. This genus is unique in the *Geminiviridae* family as it includes viruses transmitted by a treehopper vector. The topocuviruses genome is organized into six genes, a pre-coat gene, *CP*, *Ren*, *TrAP*, *Rep* and *Sd/p.sd* [12].

Typically, the "Old World Geminiviruses" (from Europe, Asia, and Africa) are predominantly monopartite and commonly associated with alpha- or betasatellite DNAs, whereas "New World Geminiviruses" (from Americas) are predominantly bipartite and may be associated with alphasatellites [13]. Betasatellites depend on their helper geminivirus for replication, systemic movement, encapsidation, and insect transmission. The betasatellite genome is approx. 1.35 kb long and harbors a single Open Reading Frame (ORF), βC1, which encodes a pathogenicity protein responsible for symptom development and suppression of RNA-mediated defenses [14]. In contrast, alphasatellite DNAs are capable of autonomous replication, but requires the helper geminiviruses for systemic movement and vector-transmission [15]. The alphasatellite genome is approx. 1.37 kb long and contains a single ORF, which encodes a rolling-circle replication-initiator protein (*Rep*) [16].

Unlike other important viral pathogens [Hepatitis C (`hcv.lanl.gov`) and HIV (`hiv.lanl.gov`)] no databases which integrate all relevant information and provides user-friendly tools enriched by machine learning approaches for easily manipulating geminivirus data, has been developed. The large amounts of information are distributed in a wide range of databases and in different file formats (for example GenBank, UniProt, VIPR and ViralZone). The access to this information is usually a complex and time-consuming task. Additionally, a high level of computational expertise is required. To address this restriction, we developed a new data warehouse designated Geminivirus Warehouse (`geminivirus.org`), using the concepts of the KDD process. The data warehouse geminivirus.org uses the ETL (Extract, Transform, Load) process, commonly applied for data warehouses, to choose, curate sequences, and standardize data. geminivirus.org is enriched with machine learning methods for classification of the viral genus using genomic sequence and identification of gene coding sequences. The computational tools comprise also species demarcation, and advanced bioinformatics tools for basic local alignment search, pairwise sequence comparison, including construction of the respective identity matrix, and an optimized phylogenetic analysis. Furthermore, we developed an algorithm for ORF prediction from the genomes of each genus with high accuracy and that is capable of identifying possible intron regions.

Construction and content

Implementation

The data warehouse geminivirus.org was implemented using an MVC (Model-View-Controller) software architecture pattern with modules programmed in the Java programming language and MYSQL server relational databases. The Data Warehouse structure was organized in SQL tables in the STAR format [17] and geminivirus.org was developed using the KDD concepts, in which the ETL process was applied and machine learning algorithms from the Weka library v3.7.11 [18] were used. The design and workflow are summarized in Figure 1 and detailed in the following sections.

Figure 1 Overview of the design at geminivirus.org. Initially, the geminivirus data were recovered from GenBank in the genbank format file (1). The data were extracted, transformed and standardized using algorithms based on rules and machine learning approaches (2). In sequence, the abstracts of the scientific publications were recovered of the PubMed (pubmed.com) (3) and the geographic coordinates of the isolates were retrieved from the Google Maps (4). The data were murged and loaded in the relational database (5) in different dimensions as collection date, host range, geographic region, genomic data, publications and organism data. The data were used to define the training set and classification models using Random Forest (RF), Multilayer Perceptron (MLP) and Sequential Minimal Optimization (SMO) algorithm (6). Information and analytical tools, such as basic local alignment tools (BLAST), sequence demarcation tools and phylogenetic reconstruction, were coupled (7) and an ORF Search tool for prediction and classification of ORFs based on machine learning approach (8) was implemented. All results of the analysis are visible and freely available without any restrictions to those outside the academic scientific community (9).

Source data

Initially, the geminivirus-related data were obtained as a document format from nucleotide Genbank database. This file format has information related to the complete genome, country of origin, region or province geographic coordinate, collection date, host, author, responsible for the collect, among others. It is important to mention that all these data are not always available and the document structure is often out of standard. Next, the articles abstracts relating to the retrieved sequences were recovered from the PubMed database. Finally, the geographic information such as country name, geographic region or geographical coordinate obtained from Genbank were used to retrieve the geographical coordinates in UTM (Universal Transverse Mercator) format on Google Maps.

Raw data extraction

Preliminarily, information was extracted from GenBank file, as described in Table 1. Then, each file record was inspected using the following criteria concerning the full-length genome sequences: i) length must be greater than or equal to the minimum predefined size, ii) length cannot exceed the maximum length, and iii) they must fit one of the seven established criteria from ICTV (www.ictvonline.org). Genomes of alpha- and betasatellites were also included, as well as genomes and proposed genera that have not yet been classified by ICTV. The Algorithm 1 describes how information was extracted and how the selection criteria were applied. Thereafter, each pre-selected record was stored as a candidate to join the data warehouse sequences.

Transformation and standardizing data

The transformation is the step that allows standardize, correct errors and relate different information or data from heterogeneous sources in order to improve the quality and consolidate the unification of data [19]. In addition, allow metadata to be associated with data of interest and subsequently entered in the database [20]. To perform this step, the pre-selected records were processed using the following criteria incorporated in Algorithm 2:

(i) **Origin of replication:** Firstly, corrections were performed in genome sequences that do not start at the same genomic coordinates. This genome region was adjusted to start in the first nucleotide after the cleavage site within the conserved nonanucleotide at the geminivirus replication origin (TAATATT▼AC;) and geminivirus-associated alpha and betasatellite DNAs;

(ii) **Repairing the Open Read Frame coordinates in the genome:** The start and stop codon coordinates of each gene belonging to the genome sequence adjusted in step I were redefined;

(iii) **Checking the consistency of ORFs:** We verified whether all coding DNA sequences had start and stop codons, and whether the amino acid sequences were not truncated;

(iv) **Standardization of gene acronyms:** The standardization of acronyms for gene identification was conducted following the genomic organization of the seven genera of the *Geminiviridae* family. The following acronyms were used: CP, capsid protein; Rep, replication-associated protein; TrAP, transactivator protein; Ren, replication enhancer; MP, movement protein; NSP, nuclear

shuttle protein; Reg, regulatory gene; Sd, symptom determinant; Ss, silencing suppressor; tgs, transcriptional gene silencing. Note that the DNA-A component of Old World bipartite begomoviruses contains a V2 ORF, defined as pre-coat in our standardization [8];

(v) **Genus classifier in the *Geminiviridae* family:** The genera were confirmed using machine learning approaches;

(vi) **ORF classifier in each genus of the *Geminiviridae* family:** In this step, we confirmed whether the ORFs were correctly standardized using machine learning approaches;

(vii) **Standardization of country abbreviation:** country and continent abbreviations for all genera were standardized;

(viii) **Standardization of species name:** The species names were replaced following a list of begomovirus species, as of January 2015 [9], available in ICTV website (talk.ictvonline.org/ictv wikis/m/files begomo/default.aspx);

(ix) **Recovering geographic coordinates:** We recovered geographical coordinates with exact (deposited coordinates) or approximate positions through secondary information, such as the informed country;

(x) **Recovering scientific publications:** All scientific publications related to a deposited sequence were recovered.

Our programs were developed using object-oriented concepts. We implemented a collection of classes designated as Object Geminivirus (OG). The instances of the OG classes have the purpose of storing information, performing tasks (e.g., create, read, update, delete data into database), and making the communication between the application (user interface) and the database. Furthermore, such objetcts allow more flexible operations in the database creation, as well as recording reading, updating, and deletion. OG objects are loaded with the data just after their transformation, preparing them to be loaded in our data warehouse.

Data load

The storage structure of our data warehouse was modeled in SQL tables with an adapted Star scheme. The Star scheme is composed by one or more fact tables that represent data as facts. For instance, each isolate or organism can have these facts: i) genome sequence and open read frames; ii) geographical localization; iii) collection date; iv) host range; vi) authors and related institution; and vii) scientific publication reference. To insert the data in SQL tables, the transformed data were loaded to OG object, by which each full-length genome and its associated metadata were inserted in the database, maintaining the integrity of data in different star tables scheme. It is worth mentioning that OG object allows the control of all changes and transformations performed in a sequence and their associated metadata, in order to record the history of changes and additional information. Other information regarding genome sequences, which was not available on the GenBank Database, was manually updated by a team, who inspected several scientific articles and posteriorly inserted the information in the data warehouse.

Data mining

Machine learning

The genome sequences and complete ORFs were reclassified using machine learning approaches. The best strategy after testing different learning models was the

Random Forest algorithm. The resulting models were able to perform genus taxonomic classification within *Geminiviridae* family with high accuracy. In addition, ORF classification models were also generated for each genus in that family. The experiments with ML algorithms that led to the selection of the Random Forest strategy are explained in more detail in the paper "Fangorn Forest (F2): A method for classification of genes and genera in the *Geminiviridae* family" (submitted).

Bioinformatics tools

The data warehouse geminivirus.org provides a user-friendly web interface for an easy usage of advanced bioinformatics tools to search for viral information and to perform basic local alignment search, species demarcation, optimized phylogenetic analysis, ORF discovery and classification, as well as geographical visualization of geminiviruses and satellite-related data.

(A) **User-friendly search modules:** The web interface contains user-friendly search modules for viral sequences and scientific publications. The user can perform a search using keywords, such as viral name, host plant, GenBank Database accession number, country of origin, genome segment (DNA-A, DNA-B, monopartite genome or alpha- and betasatellite), collection year, and sequence submission year. The search for scientific publications can also be performed using keywords such as PubMed ID, author name, virus name, scientific journal name and sequence publication year;

(B) **Basic local alignment search:** To perform a basic local alignment search with sequences of genomes, amino acids, or CDS, we embedded the BLAST software in our platform BLAST software [21] (BLASTn, BLASTp and BLASTx algorithms) with pre-adjusted p-value parameters.

(C) **Species demarcation:** We also incorporated the SDT v1.0 software [22] into geminivirus.org, which enables pairwise-sequence comparison analyses. Query sequences are used for pairwise alignments using Mafft [23], Muscle [24], or Clustalw [25] algorithms. Based on the percentage of sequence identities, the user can select desired sequences and generate a comparative identity matrix. Thus, this matrix can be viewed in geminivirus.org or downloaded to the user's computer, opened with the original SDT software and the matrix can be edited using any image editing software.

(D) **Phylogenetic reconstruction analysis:** An automatized phylogenetic analysis may be performed in geminivirus.org. The user initially enters at least one query sequence and then performs a search for sequence homology using BLAST algorithms. Query sequences are used to pairwise alignments using Mafft, Muscle, or Clustalw algorithms, and the alignment output is automatically loaded into the FastTree2 software [26]. The phylogenetic analysis is performed using the maximum-likelihood method with 1.000 bootstrap replications and other default parameters. The FastTree 2 software uses minimum-evolution subtree-pruning-regrafting and maximum-likelihood NNIs (nearest-neighbor interchange) to search for better trees. We still embedded the Phytools R package into our platform for visualization and additional analysis [27], for which the fastBM simulation function is used. In addition, the phylogenetic tree output can be also downloaded in newick format to

the user's computer, then opened and edited using, for example, the FigTree v1.4.2 software (http://tree.bio.ed.ac.uk/software/figtree).

(E) **Data Visualization:** All information related to geminiviruses and geminivirus-associated satellites, such as viral species and geographical distribution, can be visualized using a graphic interface developed in the Google Maps API (https://developers.google.com/maps/?hl=en) and Google MarkerClusterer (https://github.com/googlemaps/js-marker-clusterer). Additionally, statistical information about the amount of full-length genome sequences per country, viral species, year and related scientific publications are also shown in charts using the Google Charts API (https://developers.google.com/chart/?hl=en).

(F) **Discovery and ORF classification:** We have developed an algorithm for prediction and classification of genes. Moreover, the algorithm allows the classification of viral genus based on the genomic sequence using machine learning approaches. For the sake of organization and focus, we explain this methodology in more detail in the paper "Fangorn Forest (F2): A method for classification of genes and genera in the *Geminiviridae* family" (submitted).

Utility and discussion

Geminiviruses infect a wide range of dicotyledonous and monocotyledonous plants causing expressive losses worldwide. A wide range of studies, using genomic data and different bioinformatics tools, have been published in the literature on this subject, such as studies of molecular interaction mechanisms among viral and host plants [28][29][30], population biology [31], species taxonomy [8][9][32], and diversity discovery of new viral species [33]. In spite of the geminivirus relevance, inflicting serious threat to agriculture worldwide, there are no databases integrating all relevant related information and providing user-friendly tools for easily manipulating data. This provides motivation for creating a specific database for geminivirus and automated pipelines to facilitate research in order to boost findings and exchange information among researchers.

The high diversity and amount of viral species complicate the recovery and interpretation of viral genomic and proteomic data. After the advent of rolling-circle amplification (RCA) using the phi-29 DNA polymerase along with current high-throughput sequencing methods, thousands of full-length sequences became available from public databases in the last 10 years. This large amount of information is available in a wide range of databases or as supplemental material in scientific publications, such as full-length genome, coding DNA sequence, geographical localization, host range, data collection, species names, and species identifiers (acronym). All these information has a great potential to result in new knowledge when unified.

While 274 full-length genomic sequences of geminiviruses were available from Genbank databases from 1990 to 2003, this number increased exponentially (about 34 times) until the current year (9255 full-length sequences). A paralleled increased number of scientific papers involving geminiviruses has been reported during the same period. However, the number of full-length genomic sequences is distributed among the seven genera, 6255 begomoviruses, 1163 mastreviruses, 67 curtoviruses, 45 turncurtoviruses, 29 becurtoviruses, 6 eragroviruses, 1 topocuvirus in addition to 1154 betasatellites, 488 alphasatellites, and 47 unclassified full-length genome sequences. Recently discovered geminiviruses showed that the genetic diversity among

genera reaches high levels and, in some cases, presents specific genome architectures [8]. Despite of this highly divergent genomic content, we have built a web platform that includes associated metadata, search modules, bioinformatics tools, and machine learning methods, which retrieve information of interest, demarcate species, and classify genera and ORFs. The following sections provide details of how geminivirus.org offers a simple and user-friendly environment to retrieve or discover information about geminiviruses.

Searching for sequence and visualization of the data. The searching for geminivirus and DNA satellite and gene sequences into Geminivirus Data Warehouse is available through the menu designed Virus. The search and analysis tools provide various searching criteria on both nucleotide sequences (full-length genomes or genes) and amino acid sequences (protein). The metadata provides to the users the ability to perform searches based on parameters (or combinations thereof), such as viral name, host plant, access number into GenBank Database, origin country or genome segment (DNA-A, DNA-B begomoviruses components, alpha or betasatellite DNAs). After the search, the results are shown in a table format, where each column refers to the accession number, sequence description, collection date, submission date in GenBank, host, country and length sequence. By clicking in accession number, the user can get all information related to the complete sequence at a user-friendly screen (geminivirus.org:8080/geminivirusdw/viewOrgServlet?id=KC706589). Besides the already mentioned metadata, information, regarding the sequence authorship, authors responsible for depositing the sequences, funding institutions and responsible for data collection when available in GenBank, was preserved. Information concerning authorship is accessible in the abstract publications when linked to the complete sequences.

On the same screen, other information details the consistency and quality of data. This information is presented below:

(i) **About ORFs:** Relevant information such as gene names, chain (forward or complement), genome position, name gene, protein sequence and coding DNA sequence are also presented. In addition, an inspection of the quality of these sequences is also presented in order to assist the user regarding consistency of the sequences (presence of start codon, stop codon and truncation). These checks are illustrated using the light blue star and a note indicating the status of algorithm verification (properly verified by the algorithm). However, oversight annotation and submission of sequences are quite common. To work around this problem, ORFs were classified using the models of Random Forest ranking algorithm. Thus, the result of the classification is presented to the user, indicating the gene name and the resulting likelihood classification. This way, the consistency confirmation of ORF and its annotation is obtained.

(ii) **View genome architecture:** The genomic architectures can be viewed on an interactive circular diagram. Furthermore, the genes are shown in a table comprising the same summary information.

(iii) **Revisions:** During the process of cleaning and processing the data, some changes are performed in the complete genome, CDS or protein, whereas metadata is included in other data sources or manually entered. The added and changed information is stored in a database, in order to foment the change

history. In addition, the history is visible to the users on a time line of the change.

The treatments and aggregated information of the submitted sequences are important and positively assist in conducting several studies, such as migration studies, phylogeography analysis, recombination analysis, genetic diversity, demarcation species, among others. The associated metadata is represented by badges to show intuitively the quality and sequence reliability. The viral sequences approved in the initial filter receives an emblem illustrated as a yellow color medal. Sequences related to at least one publication receive green medal, and sequences that are inspected and corrected manually receive red color medals. In addition, those classified correctly by Randon Forest algorithm receive the blue medal. The filled stars refer to existence of a particular metadata and empty star informs the metadata absence. It is important to mention that receiving empty star does not implicate in the nonexistence of the metadata but rather it means that it is in the process of manual inspection.

Searching publications

The search for scientific publications can also be performed using keywords, such as PubMed ID, author name, virus name, scientific journal name and sequence publication year. The search results are presented in a table format containing the titles, authors, publication year and PMID number. Clicking on publications enables access to abstract of selected publication, whereas the title, related authors, scientific journal, PMID and accession number of the virus sequence associated with the selected publication can also be retrieved.

Basic local alignment search tools

The users can perform a basic local alignment search against a query sequence (nucleotide or amino acid) using BLASTn, BLASTx, BLASTp algorithms coupled into data warehouse. Moreover, the BLAST serves as a tool for initial sequences search with higher similarity. These results can be chosen and automatically used as input in the SDT v1.0 software for species demarcation and also in FastTree for phylogenetic analysis, both coupled in the data warehouse. The BLAST results are merged with other metadata associated as sequence match, collection date, host and geographic region. Thus, the tabulated results may help researchers in making decision based on sequence comparisons, host range and geographical location.

Species demarcation

The SDT (Sequence Demarcation Tool) was recently implemented for viral species demarcation and provides standardization for all parameters, such as alignments and processing gaps to calculate the percentage of sequences identity between genome or gene sequences [22]. We incorporate an adapted and parallelized version of SDT software into geminivirus.org (geminivirus.org:8080/geminivirusdw/SDTdemarcation.jsp). This enables genome sequences of geminiviruses and associated satellite DNAs to be directly compared, and eliminates the need for a pre-installed SDT version in the computer of the users. Briefly, the analysis performs a preliminary comparison of query sequence with other available sequences

into geminivirus.org using BLAST algorithms, and this enables a pre-selection of mostly related sequence. Then, SDT performs all comparisons between the sequence queries provided by the user and those that were pre-selected in the previous step of the BLAST results.

Another advantage of using SDT from geminivirus.org is that algorithm only performs comparisons involving query sequences provided by the user against those available in geminivirus.org. This reduces the analysis complexity and duration needed to generate results. Thus, analyses involving a large number of sequences may be performed within a short period of time (especially if compared with analyses performed on the user's computer). Thus, the implementation of the SDT software into geminivirus.org enables users from various platforms to use this software. Finally, the user can obtain a color array that represents the identity percentage values that can be downloaded to the user's computer as a list containing the results of all pairwise comparisons.

Phylogenetics reconstruction

The phylogenetic analysis from geminivirus.org enables a rapid visualization of phylogenetic relationships and groupings from the input sequence dataset (geminivirus.org:8080/geminivirusdw/phylogeny.jsp). Initially, the users sequence of interest is submitted against the geminivirus sequences using BLASTn algorithm. The users can select sequences based on BLAST results. The alignment is performed using MUSCLE alignment algorithm and their output is automatically loaded into FastTreeMP 2. The FastTree 2 is a tool that enables phylogeny inference for alignments with up to hundreds of sequences. It is slightly more accurate than its former version and 100-1.000-fold faster than other tools.

Prediction and classification of ORFs in full-length genome sequence. Geminiviruses contain ten different types of genes currently known. In addition, the alphasatellites encodes only alphaRep and Betasatellites, betaC1. The most common way to identify such genes is through the ORF finder tool (www.ncbi.nlm.nih.gov/projects/gorf/). However, prediction and *in silico* annotation of these ORFs require computational expertise and time to process and analyze the data. To address this restriction, we developed a method of prediction and classification of ORFs designated Fangorn Forest method, as described in the data mining session. In addition, a complete pipeline can optionally be used to classify viral genus, species demarcation, perform phylogeny and identify related species to associate them with geographical region and host range. The Fangorn Forest is freely available at geminivirus.org:8080/geminivirusdw/discoveryGeminivirus.jsp. The algorithms have been adjusted to make the prediction and classifications without the need to assign genera or species information. Only genomic sequence and identifier are assigned as input. However, parameters can be adjusted to assign the genus and cut-off for ORFs classification. The user can also choose to view proteins with unknown function. The complete pipeline options for comprehensive analysis of ORFs predictions, viral genera, phylogenetic reconstruction and the best hits of BLAST with their associated metadata (range host, geographic region). To perform these analyzes, 20 sequences are obtained following the BLAST identity criteria. The user is free to assign the amount of 30 sequences and the identity value. Thus, the

results are displayed and can be easily analyzed and the sequences can be retrieved for downloading by means of a single and user-friendly environment.

Future work and perspectives

The geminivirus.org is structured to accommodate information about geminiviruses and related DNA satellites that become available regularly. Our platform will be frequently updated with new information extracted from GenBank, scientific publications, meetings, and abstracts. The inclusion of new data sources will contribute to provide wealth of currently available information and will promote an expansion of our system to accommodate further information that can assist the interpretation of bioinformatics analysis results. Future improvements will permit further development of meta-analysis tools and natural language processing to extract knowledge about published studies and structure sequences to be deposited directly into the data warehouse. We plan to develop a mobile application to assist data collection and information exchange among researchers and geminivirus.org.

Conclusions

geminivirus.org is an integrated and open-access data warehouse that optimizes complicated and comprehensive searches that are difficult to perform using the existing tools. Therefore, it is efficient to assist in targeted search and provides accurate and concise information on all geminiviruses and geminivirus-associated satellites to the scientific community. It provides a user-friendly environment to retrieve information about: (i) the geographic distribution of geminivirus throughout the globe by an interactive map; (ii) the circular genomic structure through interactive visualization; (iii) advanced graphs with information statistics and results provided by species demarcation, phylogenetic analysis and ORF search. Its flexibility enables the addition or analysis of various taxonomy types, genome, sampling or biological data to facilitate and update information sources. Furthermore the implementation of algorithms based on machine learning approaches allows the prediction and classification of viral genes as well as the identification of the genus based on viral genomic sequences. The data sources and additional analytical tools will greatly facilitate searches in the geminirus.com information management system. geminivirus.org will not be restricted to academics and will represent a valuable resource for a broad research community; it is freely available and will represent a valuable resource for research community.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
The authors are grateful to the Furthermore National Institute of Science and Technology in Plant-Pest Interactions (INCT-IPP), Fundação de Amparo à Pesquisa do estado de Minas Gerais (FAPEMIG), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for financial support.

Author details
[1]Departamento de Informática, Universidade Federal de Viçosa, Campus Universitário, Viçosa, Brazil. [2]National Institute of Science and Technology in Plant-Pest Interactions/BIOAGRO, Campus Universitário, Vicosa, Brazil. [3]Núcleo de Biomoléculas, Universidade Federal de Viçosa, Campus Universitário, Viçosa, MG, Brazil. [4]Departamento de Bioquímica e Biologia Molecular, Universidade Federal de Viçosa, Campus Universitário, Viçosa, Brazil. [5]Departamento de Zootecnia, Universidade Federal de Viçosa, Campus Universitário, Viçosa, Brazil. [6]Departamento de Fitopatologia, Universidade Federal de Viçosa, Campus Universitário, Viçosa, Brazil.

References

1. Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S., Robinson, G.E.: Big data: astronomical or genomical? PLoS Biol **13**(7), 1002195 (2015)
2. Tsai, C.-W., Lai, C.-F., Chao, H.-C., Vasilakos, A.V.: Big data analytics: a survey. Journal of Big Data **2**(1), 1–32 (2015)
3. Dunkel, B., Soparkar, N., Szaro, J., Uthurusamy, R.: Systems for kdd: From concepts to practice. Future Generation Computer Systems **13**(2), 231–242 (1997)
4. Olshannikova, E., Ometov, A., Koucheryavy, Y., Olsson, T.: Visualizing big data with augmented and virtual reality: challenges and research agenda. Journal Of Big Data **2**(1), 1–27 (2015)
5. Ma, C., Zhang, H.H., Wang, X.: Machine learning for big data analytics in plants. Trends in plant science **19**(12), 798–808 (2014)
6. Rasheed, Z., Rangwala, H.: Metagenomic taxonomic classification using extreme learning machines. Journal of bioinformatics and computational biology **10**(05), 1250015 (2012)
7. Brown JK, B.R.Z.F.M.E..N.-C.J. Fauquet CM: Chapter family geminiviridae in virus taxonomy. 9th Report of the International Committee on Taxonomy of Viruses, 351–373 (2012)
8. Varsani, A., Navas-Castillo, J., Moriones, E., Hernández-Zepeda, C., Idris, A., Brown, J.K., Zerbini, F.M., Martin, D.P.: Establishment of three new genera in the family geminiviridae: Becurtovirus, eragrovirus and turncurtovirus. Archives of virology **159**(8), 2193–2203 (2014)
9. Brown, J.K., Zerbini, F.M., Navas-Castillo, J., Moriones, E., Ramos-Sobrinho, R., Silva, J.C., Fiallo-Olivé, E., Briddon, R.W., Hernández-Zepeda, C., Idris, A., *et al.*: Revision of begomovirus taxonomy based on pairwise sequence comparisons. Archives of virology **160**(6), 1593–1619 (2015)
10. Muhire, B., Martin, D.P., Brown, J.K., Navas-Castillo, J., Moriones, E., Zerbini, F.M., Rivera-Bustamante, R., Malathi, V., Briddon, R.W., Varsani, A.: A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus mastrevirus (family geminiviridae). Archives of Virology **158**(6), 1411–1424 (2013)
11. Soto, M.J., Gilbertson, R.L.: Distribution and rate of movement of the curtovirus beet mild curly top virus (family geminiviridae) in the beet leafhopper. Phytopathology **93**(4), 478–484 (2003)
12. BRIDDON, R.W., BEDFORD, I.D., TSAI, J.H., MARKHAM, P.G.: Analysis of the nucleotide sequence of the treehopper-transmitted geminivirus, tomato pseudo-curly top virus, suggests a recombinant origin. Virology **219**(2), 387–394 (1996)
13. Briddon, R.W., Patil, B.L., Bagewadi, B., Nawaz-ul-Rehman, M.S., Fauquet, C.M.: Distinct evolutionary histories of the dna-a and dna-b components of bipartite begomoviruses. BMC evolutionary biology **10**(1), 1 (2010)
14. Cheng, X., Wang, X., Wu, J., Briddon, R.W., Zhou, X.: βc1 encoded by tomato yellow leaf curl china betasatellite forms multimeric complexes in vitro and in vivo. Virology **409**(2), 156–162 (2011)
15. Briddon, R., Stanley, J.: Subviral agents associated with plant single-stranded dna viruses. Virology **344**(1), 198–210 (2006)
16. Fiallo-Olivé, E., Martínez-Zubiaur, Y., Moriones, E., Navas-Castillo, J.: A novel class of dna satellites associated with new world begomoviruses. Virology **426**(1), 1–6 (2012)
17. Boehnlein, M., Ulbrich-vom Ende, A.: Deriving initial data warehouse structures from the conceptual data models of the underlying operational information systems. In: Proceedings of the 2nd ACM International Workshop on Data Warehousing and OLAP, pp. 15–21 (1999). ACM
18. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD explorations newsletter **11**(1), 10–18 (2009)
19. Kumar, V., Thareja, R.: A simplified approach for quality management in data warehouse. arXiv preprint arXiv:1310.2066 (2013)
20. Bala, M., Boussaid, O., Alimazighi, Z.: Big-etl: Extracting-transforming-loading approach for big data. In: Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA), p. 462 (2015). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp)
21. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of molecular biology **215**(3), 403–410 (1990)
22. Muhire, B.M., Varsani, A., Martin, D.P.: Sdt: a virus classification tool based on pairwise sequence alignment and identity calculation. PLoS One **9**(9), 108277 (2014)
23. Katoh, K., Misawa, K., Kuma, K.-i., Miyata, T.: Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic acids research **30**(14), 3059–3066 (2002)
24. Edgar, R.C.: Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research **32**(5), 1792–1797 (2004)
25. Li, K.-B.: Clustalw-mpi: Clustalw analysis using distributed and parallel computing. Bioinformatics **19**(12), 1585–1586 (2003)
26. Price, M.N., Dehal, P.S., Arkin, A.P.: Fasttree 2–approximately maximum-likelihood trees for large alignments. PloS one **5**(3), 9490 (2010)
27. Revell, L.J.: phytools: an r package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution **3**(2), 217–223 (2012)
28. Zorzatto, C., Machado, J.P.B., Lopes, K.V., Nascimento, K.J., Pereira, W.A., Brustolini, O.J., Reis, P.A., Calil, I.P., Deguchi, M., Sachetto-Martins, G., *et al.*: Nik1-mediated translation suppression functions as a plant antiviral immunity mechanism. Nature **520**(7549), 679–682 (2015)
29. Brustolini, O.J., Machado, J.P.B., Condori-Apfata, J.A., Coco, D., Deguchi, M., Loriato, V.A., Pereira, W.A., Alfenas-Zerbini, P., Zerbini, F.M., Inoue-Nagata, A.K., *et al.*: Sustained nik-mediated antiviral signalling confers broad-spectrum tolerance to begomoviruses in cultivated plants. Plant biotechnology journal **13**(9), 1300–1311 (2015)
30. Hanley-Bowdoin, L., Bejarano, E.R., Robertson, D., Mansoor, S.: Geminiviruses: masters at redirecting and

reprogramming plant processes. Nature Reviews Microbiology **11**(11), 777–788 (2013)
31. Rocha, C.S., Castillo-Urquiza, G.P., Lima, A.T., Silva, F.N., Xavier, C.A., Hora-Júnior, B.T., Beserra-Júnior, J.E., Malta, A.W., Martin, D.P., Varsani, A., *et al.*: Brazilian begomovirus populations are highly recombinant, rapidly evolving, and segregated based on geographical location. Journal of virology **87**(10), 5784–5799 (2013)
32. Briddon, R.W., Martin, D.P., Owor, B.E., Donaldson, L., Markham, P.G., Greber, R.S., Varsani, A.: A novel species of mastrevirus (family geminiviridae) isolated from digitaria didactyla grass from australia. Archives of virology **155**(9), 1529–1534 (2010)
33. Rosario, K., Marr, C., Varsani, A., Kraberger, S., Stainton, D., Moriones, E., Polston, J.E., Breitbart, M.: Begomovirus-associated satellite dna diversity captured through vector-enabled metagenomic (vem) surveys using whiteflies (aleyrodidae). Viruses **8**(2), 36 (2016)

Tables

Table 1 Information extracted from GenBank file and stored in geminivirus.org

TAGs	Value
LOCUS	KJ939916
DEFINITION	Soybean chlorotic spot virus isolate BR:Flt14:11 segment DNA-A, complete sequence.
ORGANISM	Soybean chlorotic spot virus
PUBMED	25028472
AUTHORS	Sobrinho,R.R., Xavier,C.A.D., Pereira,H.M.B., Lima,G.S.A.,,Assuncao,I.P., Mizubuti,E.S.G., Duffy,S. and Zerbini,F.M.
JOURNAL Submitted	Departamento de Fitopatologia, BIOAGRO,,Universidade Federal de Vicosa, Av. Peter Henry Rolfs s/n, Vicosa,,Minas Gerais 36570-900, Brazil
Assembly Method	CodonCode Aligner v. 4.1.1 DEMO
Sequencing Technology	Sanger dideoxy sequencing
host	Macroptilium lathyroides
taxon	1221206
country	Brazil
segment	DNA-A
lat_lon	
collection_date	18-Mar-2012
collected_by	
CDS	199..954
gene	
note	coat protein
product	CP
protein_id	AIN36521.1
translation	MVKRDAPWRHMAGTSKVSRSSNFSPRGGGGPKNNRTSEWVNRPM . . .
ORIGIN	ACCGGATGGCCGCGCGATTTTTTATGGGCCTTATCTTTTGGGCTCGTTCTTTTGGACCGAGT GTATTTGAATTAAAGTAAAGTTATTCCCTGTCCAA . . . CCGCTATAATATT

Table 2 Terms used to name CDS in NCBI.

Genere	CDS term NCBI	Varsani Standard
Betasatellite	"beta" or "c1"	betaC1
Alphasatellite	"alpha" or "rep"	alphaRep
Begomovirus	"bv1" or "nsp" or "nuclear shuttle"	NSP
Begomovirus	"bc1" or "bc2" or "mp"	MP
All genere	"c1" or "ac1" or "rep" or al1	Rep
All genere	"c2" or "ac2" or "trap" or "al2" or "transcription activator protein"	TrAP
All genere	"c3" or "ac3" or "ren" or al3	REn
All genere	"c4" or "ac4" or al4	sd/p.sd
All genere	"c5" or "ac5"	AC5
All genere	"v1" or "av1" or "cp" or "ar1" or "capsid protein" or "coat protein"	CP
All genere	"v2" or "av2" or "pre-coat" or "precoat" or ar2	pre-coat
All genere	"v3" or "av3"	Reg

Algorithm 1 Step extraction: Extracting tags, exclusion of the incomplete sequences and redundant sequences.

Input: : Raw data in the GenBank file format

$GENBANK$ = Extract tags of each organism of the genbank file ▷ see Table algorithm1 tag details
while $GENBANK \neq \varnothing$ **do**
  $genbank \leftarrow GENBANK$
  $organism \leftarrow genbank['ORGANISM']$
  $sequence \leftarrow genbank['ORIGIN']$
  $accession \leftarrow genbank['LOCUS']$
  $length_sequence \leftarrow$ length of $sequence$
  if $accession$ not contains "NC_" **then** ▷ Remove duplicated sequence
    switch $genere$ **do**
      case $Begomovirus$
        if $length \geq (2311 - 100)$ and $length \leq (2859 + 100)$ **then**
          select $organism$
        end if
      case $Mastrevirus$
        if $length \geq (2325 - 100)$ and $length \leq (2882 + 100)$ **then**
          select $organism$
        end if
      case $Eragrovirus$
        if $length \geq (2745 - 100)$ and $length \leq (2754 + 100)$ **then**
          select $organism$
        end if
      case $Turncurtovirus$
        if $length \geq (2944 - 100)$ and $length \leq (2981 + 100)$ **then**
          select $organism$
        end if
      case $Curtovirus$
        if $length \geq (2911 - 100)$ and $length \leq (3080 + 100)$ **then**
          select $organism$
        end if
      case $Topocuvirus$
        if $length \geq (2861 - 100)$ and $length \leq (2861 + 100)$ **then**
          select $organism$
        end if
      case $Becurtovirus$
        if $length \geq (2839 - 100)$ and $length \leq (2860 + 100)$ **then**
          select $organism$
        end if
      case $Unclassified$
        if $length \geq (2383 - 100)$ and $length \leq (3208 + 100)$ **then**
          select $organism$
        end if
      case $Betasatellite$
        if $length \geq 731$ and $length \leq (1452 + 100)$ **then**
          select $organism$
        end if
      case $Alphasatellite$
        if $length \geq 955$ and $length \leq (1479 + 100)$ **then**
          select $organism$
        end if
    end if
end while

Algorithm 2 Data transformation step.

Input: : Organisms selected in the extraction stage

$SELECTED \leftarrow$ set of organisms selected in the extraction stage
while $SELECTED \neq \varnothing$ **do**
  $select \leftarrow SELECTED$
  $sequence \leftarrow select['genome']$
  if $sequence$ does not contain replication origin **then**
    $sequence \leftarrow$ corrects replication origin of $sequence$
    $select['genome'] \leftarrow sequence$
    $select['All\ ORFs'] \leftarrow$ corrects all ORFs start and stop cordon of $select['All\ ORFs']$
  end if
  $select['ML\ classification'] \leftarrow$ classifies the genera of the virus using ML
  $ORFs \leftarrow select['All\ ORFs']$
  $size \leftarrow$ quantity of $ORFs$
  while $i \leq size$ **do**
    if $ORFs[i]['sequence']$ does not contain start codon **then**
      $ORFs[i]['erro'] \leftarrow$ no start codon
    end if
    if $ORFs[i]['sequence']$ does not contain stop codon **then**
      $ORFs[i]['erro'] \leftarrow$ no stop codon
    end if
    if $ORFs[i]['sequence']$ this truncated **then**
      $ORFs[i]['erro'] \leftarrow$ truncated sequence
    end if
    $ORFs[i]['Varsani\ Standard'] \leftarrow$ standardize $ORFs[i]['cds']$ ▷ see Table algorithm2
    $ORFs[i]['ML\ classification'] \leftarrow$ classifies the type of orf using ML
    $i \leftarrow i + 1$
  end while
  $select['All\ ORFs'] \leftarrow ORFs$
  if $select['country'] \neq \varnothing$ **then**
    $select['acronym'] \leftarrow$ recover the acronym using $select['country']$
    $select['continent'] \leftarrow$ recover the continent using $select['country']$
  end if
  $select['coordenate'] \leftarrow$ search coordinates in google map
  $select['publication'] \leftarrow$ search publication in PubMed
  save $select$ in Geminivirus DW
end while

Capítulo 4

Fangorn Forest (F2): Um método para classificação de genes e gênero na família *Geminiviridae*

Artigo científico a ser submetido à revista Bioinformatics Oxford.

METODOLOGIA

Fangorn Forest (F2): Um método para classificação de genes e gênero na família Geminiviridae

Jose CF Silva[1,3†], Thales FM Carvalho[1†], Elizabeth BP Fontes[2,3] and Fabio R Cerqueira[1*]

[*]Autor correspondente: fabio.cerqueira@ufv.br
[1]Departamento de Informatica, Campus Universitário, Vicosa, Brazil
A lista completa de informações sobre os autores disponivel no final do artigo
[†]Contribuiram igualmente

Resumo

Geminivírus infectam uma extensa faixa de plantas cultivadas e não cultivadas, causando expressivas perdas econômicas no mundo todo. O estudo da diversidade de espécies, taxonomia, mecanismos de evolução, distribuição geográfica e os mecanismos de interação desses patógenos com o hospedeiro, tem aumentado expressivamente nos últimos anos. Além disso, o uso de amplificação por ciclo rolante (RCA) e avançadas abordagens de metagenômica permitiram o conhecimento do viroma e identificação de muitos agentes virais em um elevado número de espécies de plantas. Tal fato tem aumentado a complexidade de determinar a nomenclatura e classificar taxonomicamente geminivírus. Há ainda o fato do gene responsável pela replicação viral (em particular os vírus pertencentes ao gênero *Mastrevirus*) ocorrem *splicing* devido à utilização da maquinaria transcricional das células do hospedeiro. Neste sentido, as ferramentas atuais apresentam limitações na identificação de *introns*. É importante mencionar ainda os problemas desses programas quanto aos aspectos de usabilidade. Neste trabalho, propõe-se um novo método, designado Fangorn Forest (F2), baseado em abordagens de aprendizado de máquina. O F2 é capaz de classificar o gênero por uma abordagem *ab initio*, i.e., utilizando apenas a sequência genômica completa, bem como realizar predição e classificação de genes. Os modelos de aprendizado de máquina utilizados obtiveram médias superiores à 0.99 nas métricas de precisão, acurácia e sensitividade para classificação de gênero e de genes. F2 demostra ser um poderoso método para classificação de gênero da família *Geminiviridae* com alta acurácia, o que também pode ser observado na predição e classificação de genes. O método é livremente acessível no endereço geminivirus.org.

Palavras-chave: Geminivírus; aprendizado de máquina; classificação de gene; classificação de gênero; Random Forest; Multilayer Perceptron; Máquinas de Vetores de Suporte

Introdução

A família *Geminiviridae* é uma das maiores e mais significantes famílias de vírus de plantas. Nessa família estão presentes vírus que possuem genoma de DNA circular de fita simples, encapsidado em partículas icosaédricas geminadas. Geminivírus infectam inúmeras espécies de plantas cultivadas, ornamentais e plantas daninhas e causam expressivos prejuízos econômicos na agricultura e segurança alimentar no mundo. A família *Geminiviridae* é composta por sete gêneros, *Begomovirus*, *Mastrevirus*, *Becurtovirus*, *Curtovirus*, *Turncurtovirus*, *Eragrovirus* e *Topocuvirus*. Estes

gêneros são atualmente definidos por tipo de inseto vetor, gama de hospedeiro, reconstrução filogenética e organização genômica [1]. Entretanto, um oitavo gênero designado *Capulavírus* foi proposto para acomodar recém descobertas de um novo geminivírus altamente divergente[2]. Geminivírus são constituídos de um ou dois componentes genômicos (DNA-A e DNA-B), que podem ser transmitidos de forma persistente e propagativa por mosca branca (Bemisia tabaci de biotipo A ou B), em particular para begomovírus, por cigarrinhas (mastrevírus, becurtovírus e curtovírus) e *treehopper* (topocuvírus). Os gêneros *Eragrovirus* e *Turncurtovivrus* ainda não possuem vetor conhecido. Geminivírus bipartidos são compostos por DNA-A e DNA-B. Em sua maioria, são encontrados no novo mundo, ao passo que os monopartidos (composto apenas pelo DNA-A) são comumente encontrados no velho mundo.

Estudos recentes revelam a primeira ocorrência de geminivírus monopartidos (begomovírus) infectando tomate no Peru e Equador [3]. Ao contrário, begomovírus bipartidos também foram identificados no Velho Mundo (Madascascar) infectando Asystasia gangetica [4][5] e associados com a doença do mosaico em *Coccinia grandis* na Índia [6]. Em geral, as doenças causadas por geminivírus tiveram impacto econômico e social em vários continentes. Por exemplo, na Europa, ocorrem doenças do tomateiro causada por *Tomato yellow leaf curl virus* (TYLCD) [[7] e doença do trigo anão (WDVD) [8][9]. Na África, a doença do mosaico da mandioca (CMD) [10] e doenças Risca em folhas de milho (MSD) [11]. Na Ásia, doença da folha de algodão (CLCuD) [12][13] e doença do grão de bico anão (Chickpea chlorotic dwarf) e nas Américas, doença do mosaico dourado em feijão (BGMD) [14]. A organização genômica de geminivírus é altamente conservada. Entretanto, suas espécies são geneticamente divergentes, sendo que codificam de 2-7 genes, possuem região inter gênica longa, região inter gênica curta e região comum entre DNA-A e DNA-B [1]. O DNA-A codifica as proteínas CP, proteína do capsídeo; Rep, proteína associada à replicação; TrAP, proteína transativadora de genes; REn, potenciador de replicação; reg, gene regulador; sd (AC4), determinante de sintoma e AC5, recentemente estudada e funcionalmente descrita como um determinante de patogenicidade que suprime as defesas antivirais baseado em silenciamento de RNA [15]. Além disso, geminivírus monopartidos no Velho Mundo contêm uma proteína pre-coat (V2) que auxilia no movimento e transporte do genoma viral na planta.

O DNA-B (para begomovírus) é responsável pelo transporte e movimento do DNA viral na planta e codificam duas proteínas: MP (proteína de movimento) e NSP (proteína de transporte nuclear). A proteína MP atua no transporte do DNA viral de uma célula para outra e NSP no transporte intracelular do citoplasma para o núcleo. Em alguns casos, geminivírus podem associar-se com DNA betassatélites (DNA-Beta) ou alphassatélite (DNA-Alpha). Betassatélites são moléculas de DNA de aproximadamente 1.35 kb e codifica uma única ORF betaC1 (proteína de patogenicidade determinante), que atua no desenvolvimento de sintomas, modulação da gama de hospedeiros do vírus e interfere com a resposta de defesa do hospedeiro. Em contraste, alphassatélites são capazes de replicação autônoma, mas dependente de geminivírus para infecção sistêmica e do vetor de transmissão. O seu genoma contém aproximadamente 1,37 kb, e codifica uma única proteína Rep.

Pesquisas recentes têm demonstrado a alta diversidade de espécies de geminivírus, múltiplos hospedeiros e distribuição geográfica em distintas regiões do velho e novo

mundo [1][16][17][18]. Atualmente, os métodos de sequenciamento de alto rendimento de sequências, avançadas abordagens de metagenômica e diversas ferramentas de bioinformática permitiram o conhecimento do viroma e identificação de muitos agentes virais em um elevado número de espécies de plantas. Além disso, usando abordagem de amplificação por ciclo rolante (RCA) [19], milhares de sequências ou genomas completos foram amplificados, sequenciados e disponibilizados em bases de dados públicas (GenBank NCBI, geminivirus.org). No momento atual, os geminivírus são classificados baseando-se no tipo de inseto vetor, gama de hospedeiro, reconstrução filogenética e organização genômica. Isso requer conhecimento em taxonomia e em bioinformática, porque diferentes ferramentas e algoritmos podem ser utilizados. Por exemplo, para alinhar sequências utilizam-se, frequentemente, os algoritmos Muscle [20], Mafft [21], Clustalw-mpi [22] e BLAST [23] para alinhamento local de sequências. Utilizam-se, ainda, métodos como neighbour joining, máxima parcimônia, máxima verossimilhança e inferência bayesiana [24] para obter a reconstrução filogenética. Além disso, outras abordagens utilizando comparações par a par de sequências são amplamente utilizadas. Essas comparações são obtidas por meio do software SDT [25] e analisadas segundo o critério taxonômico de cada gênero. Diretrizes e protocolos são propostos para demarcação e classificação de espécies para becurtovírus, eragrovírus e turncurtovírus [1]. De maneira similar, também são propostos critérios para begomovírus [17] e mastrevírus [16]. Para avaliar a organização genômica, deve-se obter, inicialmente, as Open Reading Frames (ORFs) e suas respectivas posições no genoma. Nesta etapa, a predição de ORFs é realizada por meio da ferramenta ORF finder (www.ncbi.nlm.nih.gov/gorf/gorf.html), que embora amplamente utilizada, possui limitações na identificação de introns nessa família, anotação de sequência e usabilidade. Apesar dessas ferramentas serem robustas e auxiliarem na classificação taxonômica, são amplamente dependentes de métodos computacionais que fornecem soluções aproximadas e que se utilizam algumas vezes de bases de dados com sequências não padronizadas, não curadas e com possíveis erros de anotação. Na maioria das vezes, os métodos necessários não são integrados, o que facilitaria a automatização do processo de análise de dados e de tomada de decisão.

Aqui apresentamos um método de predição *in silico* capaz de classificar gênero e genes baseado em métodos de aprendizado de máquina (Machine Learning - ML), designado Fangorn Forest (F2). Neste método, são utilizadas apenas características genômicas comuns a qualquer genoma viral para construção de modelos de classificação. Neste estudo, em particular, foram selecionados os sete gêneros da família *Geminiviridae* e seus DNAs satélites relacionados. Neste trabalho, obteve-se oito conjuntos de treinamento, sendo um composto por características extraídas do genoma completo de geminivírus para classificar gênero e os demais compostos por características extraídas dos genes para classificar ORFs. A estes conjuntos foram aplicados três algoritmos de ML para construção dos modelos de predição: máquinas de vetores de suporte (SVM), Random Forest (RF) e Multilayer Perceptron (MLP). No processo de treinamento da abordagem SVM, utilizou-se uma implementação do algoritmo Sequential Minimal Optimization (SMO). O método Fangorn Forest consegue predizer e classificar ORFs e gêneros de vírus da família *Geminiviridae* com alta acurácia. O modelo criado com o algoritmo SMO para classificação de

gêneros alcança 0.9965 de precisão, 0.9966 de acurácia e 0.9966 de sensibilidade. Já os modelos criados com os algoritmos RF e MLP para classificação de tipos de ORFs, conseguiram, em média, precisão, acurácia e sensibilidades de 0.9992, 0.9992 e 0.9992, respectivamente. O Fanforn Forest foi acoplado como uma ferramenta de análise na data warehouse geminivirus.org.

Métodos

Fonte de dados

Inicialmente, os genomas completos, utilizados para estabelecer a demarcação de espécies, foram utilizados para compor os conjuntos de treinamento para cada gênero. No conjunto de treinamento de Begomovírus, utilizaram-se as mesmas sequências usadas na última revisão taxonômica[17]. De maneira similar, também foram selecionadas as sequências do gênero *Mastrevirus* [16], *Curtovirus* [26], *Becurtovirus*, *Eragrovirus* e *Turncurtovirus* [1]. O gênero *Topocuvirus* possui apenas uma sequência disponível em bases de dados públicas [27] e também foi incluído. Além desses, também foram selecionados os genomas completos utilizados na classificação e nomenclatura de Betassatélites [18]. Estas sequências foram obtidas na data warehouse geminivirus.org. A fim de criar um conjunto de treinamento com maior representatividade possível, também foram recuperados genomas completos de Alphasatélites e DNA-B que não possuem critérios e recomendações para demarcação. As sequências recentemente depositadas em bases de dados públicas (NCBI e geminivirus.org), também foram utilizadas para compor o conjunto de treinamento. Tais sequências contribuem para reforçar e completar o conjunto de treinamento, uma vez que estas passaram por um processo de qualidade (filtros e padronizações) no geminivirus.org, que proporciona um alto nível de confiança. Para compor o conjunto de treinamento das ORFs, foram selecionadas as respectivas ORFs contidas nos genomas utilizados nos conjuntos de treinamento citados acima.

Qualidade de dados

Inicialmente, para garantir a qualidade das sequências completas e obter conjuntos de treinamento robustos, foram estabelecidos os seguintes critérios: (i) as sequências genômicas foram padronizadas começando no sítio conservado do nono nucleotídeo da origem de replicação (TAATATTAC) [28]; (ii) cada sequência deve conter tamanho mínimo e tamanho máximo no comprimento da sequência genômica completa para cada gênero (Tabela 1); (iii) as ORFs necessariamente devem ser completas; (iv) erros de anotação de ORFs, bem como o acrônimo e suas respectivas posições de início e fim foram corrigidos; (v) as ORFs devem conter códon de iniciação, códon de terminação e não devem ser truncadas. O processo de qualidade dos dados é descrito em detalhes no artigo "Geminivirus Data Warehouse: A database enriched with machine learning approaches" (Submetido).

Extração de atributos

A família *Geminiviridae* é composta por espécies de vírus de plantas distribuídos em sete gêneros. No entanto, a organização genômica é altamente conservada entre esses gêneros. Por exemplo, o gene Rep (codificada na fita reversa) e CP (codificada

na fita senso)[1] são comuns a todos os gêneros e suas coordenadas no genoma são aproximadamente equivalentes em relação a sua origem de replicação. Embora exista uma alta conservação na estrutura genômica, características comuns a qualquer genoma viral foram selecionadas, a fim de construir modelos de ML capazes de distinguir os gêneros e suas respectivas ORFs.

Os atributos selecionados para construir o modelo de classificação de gêneros diversificam-se desde o tamanho da sequência até as proporções de ácidos nucleicos. Por meio de uma sequência genômica completa, calculou-se o comprimento da sequência e as proporções de ademina (A), timina (T), citosina (C) e guanima (G). Posteriormente, a sequência genômica foi dividida em quatro regiões iguais (R1, R2, R3 e R4) e, para cada uma dessas regiões, calculou-se as proporções de A, T, C e G e o percentual de conteúdo GC. Dessa forma, o modelo de classificação constituiu-se de 25 atributos, que são apresentados na Tabela suplementar 1.

Para construir modelos de classificação de genes, obtiveram-se os atributos por meio da sequência de DNA codificante (CDS) e a sequência de aminoácidos. Primeiramente, extraiu-se da sequência de CDS completa, atributos como a orientação da ORF no genoma (forward/complement), o tamanho da sequência, a proporção de nucleotídeos da CDS em relação ao genoma completo e calculou-se a porcentagem de A, T, C e G dessa CDS. Além disso, a sequência CDS foi dividida em duas regiões iguais e para cada uma dessas regiões as proporções de A, T, C e G também foram calculadas. Além desses atributos, na sequência de aminoácidos obtive-se a proporção de cada um dos 20 aminoácidos primários da sequência. Neste sentido, construiu-se para cada gênero um modelo de classificação específico, contendo 35 atributos, que são apresentados na Tabela suplementar 2.

Avaliação de atributos

Avaliar as características das sequências genômicas é importante, pois permite identificar quais delas ajudam a distinguir um gênero do outro no processo de classifcação. Da mesma maneira, mensurar a relevância dos atributos de sequências de DNA codificantes (CDS) e aminoácidos permite verificar como contribuem para classificação de genes.

Assim, para avaliar a importância de cada atributo nos conjuntos de treinamento, utilizou-se o método Information Gain (IG) [29]. Tal método é baseado em entropia da informação, que permite avaliar atributos medindo o ganho de informação em relação às classes. Neste caso, cada gênero representa uma classe no conjunto de treinamento do modelo de classificação de gêneros. No conjunto de treinamento do modelo de classificação de genes, cada tipo de gene representa uma classe. O método IG é definido por $IG(Atributo) = Entropia(Classe) - Entropia(Classe|Atributo)$, em que a entropia é dada por $-\sum p_i log_2 p_i$,onde p_i é a probabilidade da classe.

Algoritmo de predição de gene

Para realizar a predição de genes, implementou-se um algoritmo baseado em abordagem gulosa como parte do método F2, designado aqui como Millau Bridge (MB). O algoritmo, inicialmente, identifica todos os possíveis códons de iniciação [ATG (5' $\rightarrow$ 3') e CAT (3' $\rightarrow$ 5')] e a fase de leitura na sequência senso ou anti-senso. Além disso, determina as coordenadas onde se localizam o códon de iniciação e

códon de terminação no genoma. Da mesma forma, todos os códons de terminação [TAA, TAG,TGA (5' → 3') e TTA, CTA,TCA(3' → 5')] são localizados. Cada códon de iniciação da sequência no sentido 5' → 3' é pareado com os códons de terminação no mesmo sentido 5' → 3'. Na sequência, restrições são estabelecidas para checar a consistência das ORFs: (i) se as sequências estão em fase de leitura e se o comprimento é múltiplo de 3; (ii) caso a condição (i) tenha sido satisfeita, a sequência CDS candidata é traduzida em sequência de aminoácidos; (iii) verifica-se se a sequência de aminoácidos está truncada ou se possui tamanho igual ou superior a 33 aminoácidos. Similarmente, essas regras são aplicadas para os códons de iniciação e terminação localizados no sentido 3' → 5'.

Entretanto, na família *Geminiviridae* foram reportados genes que codificam diferentes formas de splicing na orientação 3' → 5' de sequências genômicas de maize streak virus (MSV) [30]. Neste sentido, regras básicas do processo biológico de excisão de mRNAs [31] foram utilizadas, a fim de identificar regiões de splicing com alta precisão. Nesta etapa, os códons de iniciação e terminação podem ou não encontram-se na mesma fase de leitura no sentido 3' → 5'. Em resumo, dando continuidade aos passos descritos acima: (iv) percorrem-se todos os stop códons no sentido 3' → 5'; e verifica-se (v) se a posição do códon de terminação é maior que a posição do códon de iniciação. Em seguida, (vi) os sítios de excisão (CT e AC) são localizados; (vii) cada sítio de excisão CT candidato é pareado com todos os possíveis sítios AC, então são verificados (viii) os tamanhos dos éxons (éxons 1 min. 204 e éxons 2 min. 148 nt) e do íntron (min. 67 nt, max. 102 nt), respectivamente. Depois (ix) é verificado se a quantidade de pirimidinas é maior que a de purinas aos 50 pb "upstream" aos sítios de excisão AC; (x) o tamanho mínimo (1000 nt) é verificado e se as sequências estão na fase correta de leitura (se comprimento possui tamanho múltiplo de 3); (xi) obtém-se o reverso complementar da sequência; (xii) a sequência CDS candidata é traduzida se não estiver truncada. As restrições dos tamanhos dos éxons, íntrons e tamanhos das sequências foram determinadas observando as estruturas dos genes dessa família, em particular Mastrevírus que possui íntron e faz a junção de éxons denominando o gene C1:C2.

Escolha do algoritmo de aprendizado de máquina

O método Fangorn Forest é constituído de oito modelos de ML, cada um proveniente de um conjunto de treinamento específico (Tabela 1). O modelo 1 (GM1) foi ajustado para classificar genomas completos dos sete gêneros da família *Geminiviridae* e DNAs satélites relacionados, utilizando 25 atributos. Os demais modelos foram treinados para classificar os genes de cada gênero, sendo que o modelo 2 (BM2) foi ajustado para classificar os genes de begomovírus; o modelo 3 (MM3) ajustado para classificar ORFs de mastrevírus; o modelo 4 (BM4) para classificar os genes de becurtovírus; o modelo 5 (CM5) para classificar os genes de curtovírus; o modelo 6 (TM6) para classificar genes de turncurtovírus; o modelo 7 (EM7) para classificar genes de eragrovírus e o modelo 8 (ABM8) foi ajustado para classificar genes de alphasatélite e betasatélites.

Neste trabalho foram testados três algoritmos de ML, a fim de encontrar o algoritmo que melhor se ajusta a cada conjunto de treinamento. Os três algoritmos de ML testados foram: Sequential Minimal Optimization (SMO), Random Forest (RF)

e Multilayer Perceptron (MLP). Tais algoritmos são implementados na suite Weka v3.7.11. O algoritmo SMO é um método frequentemente utilizado para a resolução do problema de programação quadrática no qual a abordagem SVM se baseia para encontrar o hiperplano de margem máxima para separação de duas classes [32]. Os experimentos realizados com o SMO foram feitos com a API do Weka em códigos na linguagem Java. Quatro parâmetros foram explorados no caso do SMO: (i) Kernel: funções que realizam o mapeamento dos dados num espaço de maior dimensão para permitir uma melhor separação de dados que não são linearmente separáveis; (ii) C: chamado de parâmetro de complexidade, permite o ajuste das margens do hiperplano separador. Quanto maior o valor de C, mais estreita será a margem, o que significará uma menor tolerância a classificações errôneas do modelo com relação aos dados de treinamento; (iii) Cache Size: ajuste no tamanho do cache do kernel; (iv) Exponent: valor do expoente da função kernel Polykernel.

O algoritmo RF é um método de classificação baseado em árvores de decisão, capaz de fazer regressão e classificação. A classificação de uma nova instância se dá pela classificação de múltiplas árvores, resultando em um consenso dessas classificações por meio de um procedimento de votação (emsemble) [33]. Para obter o melhor modelo, os parâmetros foram estimados alterando (i) o número de árvores (Num Trees), (ii) a quantidade de atributos para ser usado na seleção aleatória destes (Num Features), e (iii) a semente para geração de números aleatórios (Seed).

O algoritmo MLP é um tipo de rede neural amplamente utilizado devido ao seu alto poder de predição em sistemas não-lineares, em que vários estudos apontam os benefícios que as redes neurais oferecem quando comparados com tradicionais técnicas de modelagem estatística[34]. O MLP é composto por três tipos de camadas de neurônios artificiais. O número de atributos do conjunto de treinamento define a quantidade de neurônios da camada de entrada. Os valores de cada atributo são atribuídos a essa camada para que sejam processados pela rede. À frente da camada de entrada, e antes da de saída, pode haver uma ou mais camadas intermediárias, constituídas de um ou mais neurônios. Cada neurônio dessas camadas podem conectar-se apenas aos neurônios da camada subsequente (feedforward connections). Essas conexões possuem pesos que definem como os valores dados de entrada serão processados para gerar a saída final. Os pesos são os valores definidos no processo de treinamento, de modo que a saída da rede tenha a maior congruência possível com a saída esperada (observada no conjunto de treinamento). Por fim, a camada de saída fornecerá o valor final processado. A retropropagação (*backpropagation*) é o método mais popular de aprendizagem (i.e., ajuste dos pesos) de uma MLP. Neste caso, as diferenças no valor produzido da rede com relação ao valor observado no conjunto de treinamento são utilizadas para ajuste dos pesos das conexões no sentido oposto ao da computação da rede (i.e., da camada de saída para a camada de entrada). Nos experimentos realizados com esse algoritmo, só utilizou-se uma camada intermediária (também chamada de oculta). Os seguintes parâmetros foram explorados: (i) quantidade de nodos na camada oculta, (ii) valor da taxa de aprendizado, (iii) valor de momentum e (iv) número máximo de iterações (epochs).

Nos experimentos, cada conjunto de treinamento deu origem a 49.491 diferentes modelos de ML. Sendo que 2.560 modelos foram obtidos com SMO, 40.931 modelos

com RF e 6.000 modelos com MLP. Cada um desses originou-se de um desses algoritmos com diferentes combinações de parâmetros (Tabela suplementar 3). Para cada execução, avaliram-se os modelos por meio da abordagem de validação cruzada com 10 partições [35] e obtiveram-se as seguintes métricas de avaliação da performance dos modelos: acurácia (Equação 1), taxa de falso positivo (Equação 2), precisão (Equação 3), sensibilidade (Equação 4) e F-Measure (Equação 5). Após a execução dos experimentos, foram escolhidos os algoritmos e os melhores parâmetros que resultaram em maior precisão e acurácia nos modelos de ML em cada conjunto de treinamento. Portanto, oito modelos de classifcação foram determinados para compor o método Fangorn Forest.

$$Acurácia = \frac{TP + TN}{P + N} \tag{1}$$

$$Taxa\ de\ Falsos\ Positivos\ (FPR) = \frac{FP}{FP + TN} = 1 - TNR \tag{2}$$

$$Precisão = \frac{TP}{TP + FP} \tag{3}$$

$$Sensibilidade\ ou\ Taxa\ de\ Verdadeiros\ Positivos\ (TPR) = \frac{TP}{TP + FN} \tag{4}$$

$$F-measure = \frac{2TP}{2TP + FP + FN} \tag{5}$$

Método Fangorn Forest

O método Fangorn Forest foi estabelecido em 3 etapas distintas, em que cada etapa permite ao método tomar decisões de forma autônoma, baseando-se em informações adquiridas na classificação. Além disso, automaticamente uma quarta etapa pode ser efetuada, para que outras ferramentas de análises tradicionais sejam utilizadas. Tais etapas são ilustradas na Figura 1. Inicialmente o método F2 utiliza o modelo GM1 para classificar um genoma completo dentre os 7 gêneros da família *Geminiviridae* ou DNAs satellites relacionados (alphassatélite ou betassatélite). Na sequência, o método utiliza o algoritmo MB para predizer ORFs contidas no genoma completo. Por último, o método seleciona automaticamente o modelo de ML (BM2, MM3, BM4, CM5, TM6, EM7 ou ABM8), baseando-se na classificação da primeira etapa. Posteriormente, cada ORF predita na segunda etapa é submetida ao modelo selecionado. Tal modelo classifica a ORF como: pre-coat, Reg, CP, AC5, REn, TrAP, Rep ou sd/p.sd para DNA-A (mono ou bipartido). Em caso de DNA-B, classifica como NSP ou MP. Os genes betaC1 e alphaRep são classificados para alpha e betassatélite, respectivamente. Além disso, a organização genômica pode ser interativamente visualizada de maneira similar às estruturas sugeridas por Varsani e colaboradores [1].

Opcionalmente, o método permite análises utilizando a sequência genômica completa. Baseando-se no resultado da classificação de gênero, automaticamente são realizadas análises em pipeline, tais como: (i) BLASTn com e-value de $1.0x10^{-5}$, como objetivo de identificar as espécies mais próximas; (ii) Reconstrução filogenética

(BLASTn com e-value $1.0x10^{-5}$, alinhamento de sequências com Muscle, construção da árvore com FastTree e pacote phytools para visualização da árvore); (iii) Demarcação de espécie usando SDT software.

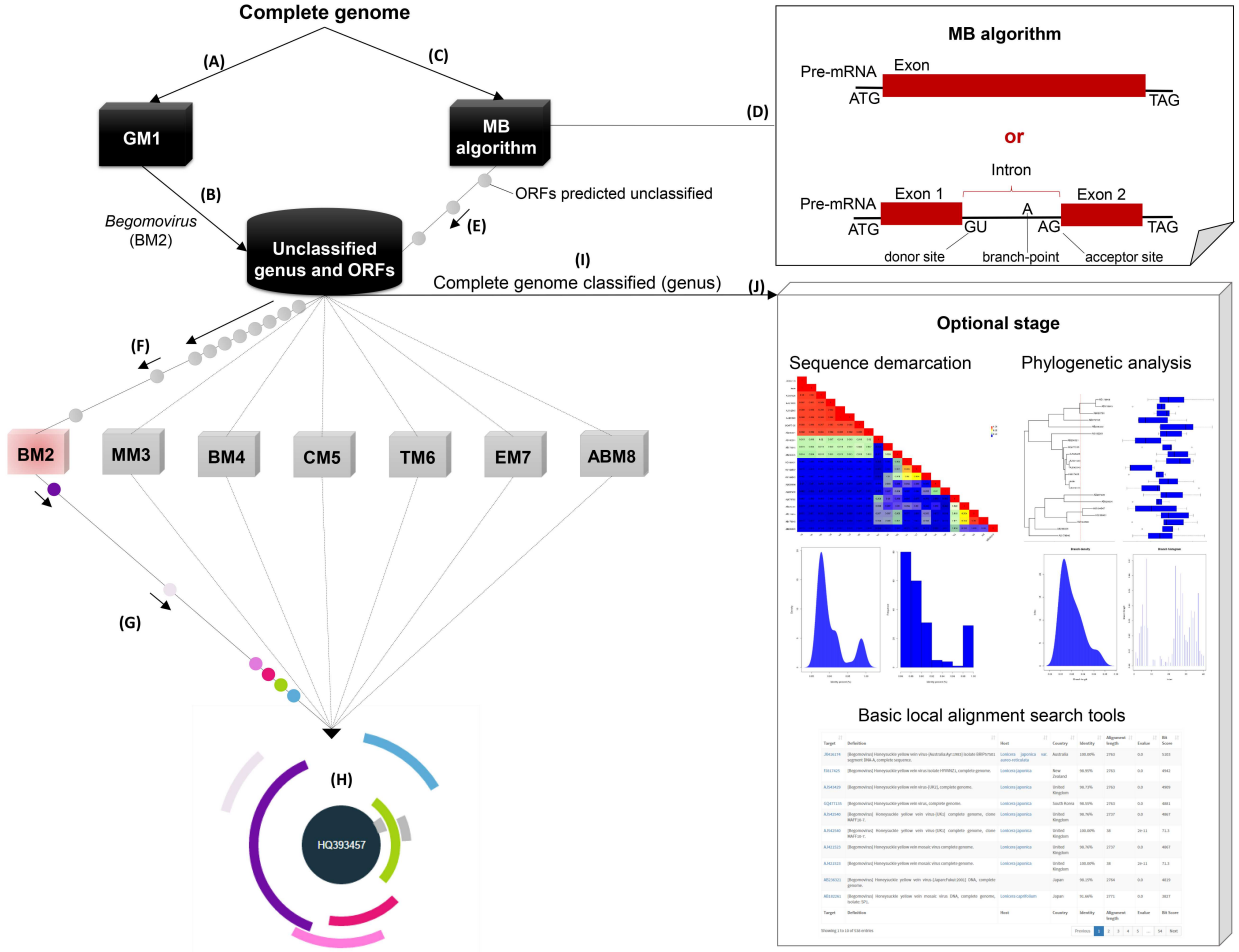

Figura 1 Fluxograma do método Fangorn Forest O genoma completo é submetido ao modelo GM1 (modelo de classificação de gêneros) (A). O resultado da classificação indica qual modelo de classificação de genes será utilizado (B). O genoma completo é submetido ao algoritmo MB. O algoritmo realiza a predição de genes (ORFs) procurando por padrões de excisão de íntrons (D). Os genes são rotulados como não classificados e direcionados para o estágio de preparação para classificação (E). Os genes não classificados são submetidos ao modelo BM2 (modelo de classificação de genes para o gênero Begomovirus)(F), enquanto os demais (MM3 (Mastrevitus), BM4 (Becurtovirus), CM5 (Curtovirus), TM6 (Turncurtovirus), EM7 (Eragrovirus), ABM8 (Alpha- e Betasatellites)) não utilizados porque não foram classificados pelo modelo GM1. Os genes classificados são encaminhados para visualização da organização genômica (G). Os genes classificados formam a organização genômica e são apresentados em formato circular (H). Baseando-se na classificação do modelo GM1, opcionalmente são realizadas análises com o algoritmo BLAST para selecionar sequências específicas do gênero classificado. Análises de demarcação de espécies (SDT) e análises filogenéticas também podem ser realizadas.

Resultado e discussão

O número de trabalhos científicos relacionados à família *Geminiviridae* tem crescido expressivamente nos últimos dez anos (geminivirus.org:8080/geminivirusdw/statistics.jsp). A alta diversidade de espécies [17] [16] [18], o elevado número de sequências completas e descobertas de novos geminivírus [2] [36] [37] [38], têm aumentado a complexidade de determinar a nomenclatura e classificar taxonomicamente geminivírus. Além disso, em algumas espécies do gênero *Mastrevirus*, transcritos primários de genes podem ocorrer modificações pós-transcricionais. Isso foi demonstrado para *maize streak virus*, que codifica o gene C1:C2 [30]. Processamento pós-transcripcional de genes é comum em eucariotos, raro em procariotos, e ocorre por meio de uma série de reações catalisadas pelo *spliceosome* ou mecanismos de auto-*splicing* [39]. As ferramentas tradicionais de predição de ORFs, por exemplo, a ferramenta ORF *Finder*, usadas por pesquisadores de geminivírus não foram adaptadas para a possibilidade de *splicing*. Para suprir essa necessidade, neste trabalho desenvolveu-se o algoritmo MB como parte do método, que utiliza-se de regras básicas de identificação dos sítios de excisão de íntrons para predição ORFs com *splicing* em geminivírus. Para tal, a aplicação de uma abordagem gulosa permitiu comparar se todas as possíveis combinações de códons de iniciação e terminação formariam ORFs completas com ou sem íntrons. A ideia do algoritmo é começar pelas extremidades das ORFs com os códons de iniciação e terminação, verificando cada possibilidade de formar uma ORF completa.

Apesar do contraste entre a alta divergência das sequências genômicas e a conservação da organização genômica possibilitar a separação entre gêneros (devido a ausência ou baixa homologia de sequência), o tamanho do genoma e as proporções de ácidos nucleico demonstram-se como atributos significativos para distinguir os gêneros dessa família. Na avaliação do ranking de atributos através do método IG, o comprimento do genoma demostrou-se a característica que mais contribui para que algoritmos possam distinguir um gênero do outro. Em segundo lugar, observou-se que as proporções de citosina (C) das regiões R1, R2, R3, são igualmente relevantes. Além desses atributos, as proporções de adenina (A) das regiões R1, R2 e R3 e a porcentagem de conteúdo GC das regiões R1, R2 e R3 também demonstraram importância significativa para a generalização dos modelos de treinamento. Em resumo, observou-se que todos os atributos escolhidos para treinar os modelos de classificação de gênero, demonstraram importância porque obtiveram valores próximos no ranking de atributos (Tabela suplementar 1). Na avaliação dos atributos dos conjuntos de treinamento, os atributos obtidos das sequências de aminoácidos obtiveram os melhores rankings na avaliação de atributos comparados com os aqueles extraídos das ORFs, embora o comprimento da ORF em sua maioria, tenha obtido as primeiras colocações (Tabela suplementar 2). Ao analisar o ranking de atributos do conjunto de treinamento MB2, observou-se que os três atributos melhores rankeados foram a proporção de Tirosina (IG=2.3807), o comprimento da ORF (IG=2.3724) e Fenilalanina (IG=2.3292), respectivamente. Os atributos melhores ranqueados para o conjunto de dados de MM3 foram Histidina (IG=1.3875), a orientação da ORF no genoma (complementar ou reverso - *chain*), e comprimento da ORF. No conjunto de BM4, os atributos comprimento da ORF (IG=2.32193), Tirosina (IG=2.32193) e Isoleucina (IG=2.32193)

foram os melhores. No conjunto de treinamento CM5, o atributo comprimento da ORF também foi o melhor ranqueado, seguido da porcentagem que cada gene representa do genoma (*percent genome*: IG= 2.58167) e proporção de Triptofano (IG=2.52714). No conjunto de treinamento TM6, os mais bem avaliados foram comprimento da ORF (IG=2.5850), proporção de Ácido Aspártico (IG=2.5850) e proporção de Glutamina (IG=2.5850). Diferentemente dos demais, no conjunto de treinamento EM7 (Eragrovirus), os atributos mais destacados foram proporção de Valina (IG=1.585), Alanina(IG=1.585) e percentual de citosina da região R2 (IG=1.585). Já os atributos que melhor distinguem alha e betassatélites são comprimento da ORF (IG=0.911), porcentagem do comprimento da ORF em relação ao genoma (IG=0.91) e proporção de Isoleocina (IG=0.883) respectivamente.

Posteriormente à avaliação desses atributos, cada conjunto de treinamento foi submetido aos três algoritmos de ML (SMO, RF e MLP). Cada algoritmo foi treinado por meio de diversas combinações de parâmetros. Os algoritmos que obtiveram os melhores valores de precisão e acurácia foram selecionados para criação dos modelos de classificação utilizados no F2 (Tabela 2). Dentre os 49.491 experimentos realizados (variando algoritmos e parâmetros) com o conjunto de treinamento contendo cada gênero da família *Geminiviridae*, o que obteve maior precisão e acurácia foi o algoritmo SMO. Por meio de validação cruzada com 10 partições, esse modelo obteve resultados de 0.9965 de precisão, 0.9966 de acurácia e 0.9966 de sensibilidade (Tabela 3). Entende-se que esses resultados permitem observar que a alta divergência de sequências entre os gêneros favorecem ao modelo a distinguir os gêneros com alta precisão utilizando as características extraídas. Além disso, a avaliação da importância dos atributos permitiu identificar quais regiões genômicas e atributos favorecem essa separação. Utilizando a mesma abordagem anterior, para criar os modelos BM2, MM3, BM4, CM5, EM7 e ABM8, o algoritmo que melhor se ajustou foi o RF. O modelo BM2 criado com RF obteve precisão, acurácia e sensibilidade de 0.99459, 0.99457 e 0.99457, respectivamente. Nos demais modelos, a performance do algoritmo RF obteve 100% de desempenho nas métricas precisão, acurácia e sensibilidade. O modelo TM6 foi criado utilizando o algoritmo MLP que também obteve 100% de desempenho nas métricas precisão, acurácia e sensibilidade. Os resultados obtidos confirmam que os atributos selecionados são apropriados para distinguir os genes em cada gênero. Em adição, as características ranqueadas para cada conjunto de treinamento mostram que embora os gêneros possuam diferentes atributos, os algoritmos conseguem generalizar facilmente o processo de classificação de gene. Além disso, a baixa homologia entre os genes facilita a distinção destes pelos algoritmos.

A combinação do modelo de classificação GM1, o algoritmo MB para predição de ORFs e os modelos de classificação de ORFs dão origem a um rápido e eficiente método. Os resultados do método F2 mostram um excelente desempenho na identificação do gênero, bem como a predição e a anotação completa das sequências genômicas desse gênero. Como complemento a esses acurados processos de classificação, o método pode opcionalmente usar a ferramenta de demarcação de espécies SDT e realizar análises filogenéticas. Para isso, o F2 está adaptado para atuar de maneira autônoma, baseando-se na classificação do modelo GM1. O resultado obtido no modelo GM1 direciona essas análises para bases de dados específicas de cada gênero. Assim, um conjunto de sequências é selecionado e utilizado para realizar essas análises.

Conclusão

Geminivírus é uma das importantes famílias de vírus de plantas. A identificação de gêneros dessa família exige cautela, pois tem se tornado um desafio devido ao elevado número de sequências disponíveis em bases de dados. Além disso, um conhecimento alto em taxonomia e análises de bioinformática é atualmente requerido.

Como resultado dessa pesquisa, um novo método composto por técnicas de aprendizado de máquina, designado Fangorn Forest, é sugerido. Tal método é estabelecido em três etapas distintas, em que, inicialmente, é capaz de classificar um genoma completo dentro de um dos sete gêneros da família *Geminividae* ou DNAs *satélites* relacionados (Alphasatellite ou Betassatélite). Posteriormente, um robusto algoritmo de predição de genes (ORFs) também é sugerido como parte desse método, designado Millau Bridge (MB). Esses genes preditos são classificados por modelos de classificação acurados, específicos para cada gênero.

Fangorn Forest se diverge das abordagens tradicionais, uma vez que não se baseia em abordagens filogenéticas, homologia de sequências ou comparação par a par. Os resultados obtidos nesta pesquisa demonstram que as características escolhidas contribuíram de forma significativa para a generalização dos modelos de classificação. Em resumo, Fangorn Forest demostra ser um eficiente método para classificação de gênero da família *Geminiviridae* com alta precisão e eficiente na predição e classificação de genes. Além disso, o método F2 pode realizar análises complementares a fim de dar suporte à tomada de decisão do pesquisador. O método é livremente acessível no endereço geminivirus.org.

Agradecimentos

Os autores agradecem a Universidade Federal de Viçosa (UFV), Instituto Nacional de Ciência e Tecnologia em Interações Planta-Praga (INCTIPP), Departamento de Informática - UFV e Divisão de Suporte ao Desenvolvimento Científico e Tecnológico - UFV. Agradecimentos pelo apoio financeiro: CAPES, CNPq, FAPEMIG and INCTIPP.

Detalhes dos autores

[1]Departamento de Informatica, Campus Universitário, Vicosa, Brazil. [2]Departamento de Bioquimica, Campus Universitário, , Vicosa, Brazil. [3]Instituto Nacional de Ciência e Tecnologia em Interações Planta-Praga/BIOAGRO, Campus Universitário, , Vicosa, Brazil.

Referências

1. Varsani, A., Navas-Castillo, J., Moriones, E., Hernández-Zepeda, C., Idris, A., Brown, J.K., Zerbini, F.M., Martin, D.P.: Establishment of three new genera in the family geminiviridae: Becurtovirus, eragrovirus and turncurtovirus. Archives of virology **159**(8), 2193–2203 (2014)
2. Roumagnac, P., Granier, M., Bernardo, P., Deshoux, M., Ferdinand, R., Galzi, S., Fernandez, E., Julian, C., Abt, I., Filloux, D., *et al.*: Alfalfa leaf curl virus: An aphid-transmitted geminivirus. Journal of virology **89**(18), 9683–9688 (2015)
3. Melgarejo, T.A., Kon, T., Rojas, M.R., Paz-Carrasco, L., Zerbini, F.M., Gilbertson, R.L.: Characterization of a new world monopartite begomovirus causing leaf curl disease of tomato in ecuador and peru reveals a new direction in geminivirus evolution. Journal of virology **87**(10), 5397–5413 (2013)
4. Parrella, G., Scassillo, L., Crescenzi, A., Nappo, A.: Typing of tomato yellow leaf curl viruses and their vector in italy. Communications in agricultural and applied biological sciences **71**(3 Pt B), 1229–1236 (2005)
5. De Bruyn, A., Harimalala, M., Hoareau, M., Ranomenjanahary, S., Reynaud, B., Lefeuvre, P., Lett, J.-M.: Asystasia mosaic madagascar virus: a novel bipartite begomovirus infecting the weed asystasia gangetica in madagascar. Archives of virology **160**(6), 1589–1591 (2015)
6. Nagendran, K., Satya, V., Mohankumar, S., Karthikeyan, G.: Molecular characterization of a distinct bipartite begomovirus species infecting ivy gourd (coccinia grandis l.) in tamil nadu, india. Virus genes, 1–6 (2016)
7. Nannini, M., Foddi, F., Murgia, G., Pisci, R., Sanna, F., Testa, M., Accotto, G.: An epidemiological survey of TYLCD in southern sardinia (italy). Communications in agricultural and applied biological sciences **74**(3), 831–841 (2008)
8. Manurung, B., Witsack, W., Mehner, S., Grüntzig, M., Fuchs, E.: The epidemiology of wheat dwarf virus in relation to occurrence of the leafhopper psammotettix alienus in middle-germany/binari manurung. The epidemiology of Wheat dwarf virus in relation to occurrence of the leafhopper Psammotettix alienus in Middle-Germany/Binari Manurung (2004)
9. Schubert, J., Habekuß, A., Wu, B., Thieme, T., Wang, X.: Analysis of complete genomes of isolates of the wheat dwarf virus from new geographical locations and descriptions of their defective forms. Virus genes **48**(1), 133–139 (2014)

10. Patil, B.L., Fauquet, C.M.: Cassava mosaic geminiviruses: actual knowledge and perspectives. Molecular Plant Pathology **10**(5), 685–701 (2009)
11. Shepherd, D.N., Martin, D.P., van der Walt, E., Dent, K., Varsani, A., Rybicki, E.P.: Maize streak virus: an old and complex 'emerging'pathogen. Molecular plant pathology **11**(1), 1–12 (2010)
12. Briddon, R.W., Markham, P.: Cotton leaf curl virus disease. Virus research **71**(1), 151–159 (2000)
13. Sattar, M.N., Kvarnheden, A., Saeed, M., Briddon, R.W.: Cotton leaf curl disease–an emerging threat to cotton production worldwide. Journal of General Virology **94**(4), 695–710 (2013)
14. Faria, J.C., Maxwell, D.P.: Variability in geminivirus isolates associated with phaseolus spp. in brazil. Phytopathology **89**(3), 262–268 (1999)
15. Li, F., Xu, X., Huang, C., Gu, Z., Cao, L., Hu, T., Ding, M., Li, Z., Zhou, X.: The ac5 protein encoded by mungbean yellow mosaic india virus is a pathogenicity determinant that suppresses rna silencing-based antiviral defenses. New Phytologist **208**(2), 555–569 (2015)
16. Muhire, B., Martin, D.P., Brown, J.K., Navas-Castillo, J., Moriones, E., Zerbini, F.M., Rivera-Bustamante, R., Malathi, V., Briddon, R.W., Varsani, A.: A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus mastrevirus (family geminiviridae). Archives of Virology **158**(6), 1411–1424 (2013)
17. Brown, J.K., Zerbini, F.M., Navas-Castillo, J., Moriones, E., Ramos-Sobrinho, R., Silva, J.C., Fiallo-Olivé, E., Briddon, R.W., Hernández-Zepeda, C., Idris, A., *et al.*: Revision of begomovirus taxonomy based on pairwise sequence comparisons. Archives of virology **160**(6), 1593–1619 (2015)
18. Briddon, R., Brown, J., Moriones, E., Stanley, J., Zerbini, M., Zhou, X., Fauquet, C.: Recommendations for the classification and nomenclature of the dna-β satellites of begomoviruses. Archives of virology **153**(4), 763–781 (2008)
19. Inoue-Nagata, A.K., Albuquerque, L.C., Rocha, W.B., Nagata, T.: A simple method for cloning the complete begomovirus genome using the bacteriophage φ29 dna polymerase. Journal of virological methods **116**(2), 209–211 (2004)
20. Edgar, R.C.: Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research **32**(5), 1792–1797 (2004)
21. Katoh, K., Standley, D.M.: Mafft multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution **30**(4), 772–780 (2013)
22. Li, K.-B.: Clustalw-mpi: Clustalw analysis using distributed and parallel computing. Bioinformatics **19**(12), 1585–1586 (2003)
23. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic acids research **25**(17), 3389–3402 (1997)
24. Yang, Z., Rannala, B.: Molecular phylogenetics: principles and practice. Nature Reviews Genetics **13**(5), 303–314 (2012)
25. Muhire, B.M., Varsani, A., Martin, D.P.: Sdt: a virus classification tool based on pairwise sequence alignment and identity calculation. PLoS One **9**(9), 108277 (2014)
26. Varsani, A., Martin, D.P., Navas-Castillo, J., Moriones, E., Hernández-Zepeda, C., Idris, A., Zerbini, F.M., Brown, J.K.: Revisiting the classification of curtoviruses based on genome-wide pairwise identity. Archives of virology **159**(7), 1873–1882 (2014)
27. BRIDDON, R.W., BEDFORD, I.D., TSAI, J.H., MARKHAM, P.G.: Analysis of the nucleotide sequence of the treehopper-transmitted geminivirus, tomato pseudo-curly top virus, suggests a recombinant origin. Virology **219**(2), 387–394 (1996)
28. Arguello-Astorga, G., Lopez-Ochoa, L., Kong, L.-J., Orozco, B.M., Settlage, S.B., Hanley-Bowdoin, L.: A novel motif in geminivirus replication proteins interacts with the plant retinoblastoma-related protein. Journal of virology **78**(9), 4817–4826 (2004)
29. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: ICML, vol. 97, pp. 412–420 (1997)
30. Wright, E., Heckel, T., Groenendijk, J., Davies, J., Boulton, M.: Splicing features in maize streak virus virion-and complementary-sense gene expression. The plant journal **12**(6), 1285–1297 (1997)
31. Nelson, D.L., Lehninger, A.L., Cox, M.M.: Lehninger Principles of Biochemistry. Macmillan, New York (2008)
32. Platt, J.C.: 12 fast training of support vector machines using sequential minimal optimization. Advances in kernel methods, 185–208 (1999)
33. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
34. Gardner, M.W., Dorling, S.: Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric environment **32**(14), 2627–2636 (1998)
35. Kohavi, R., *et al.*: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai, vol. 14, pp. 1137–1145 (1995)
36. Lim, S., Igori, D., Zhao, F., Moon, J.S., Cho, I.S., Choi, G.S.: First report of grapevine red blotch-associated virus on grapevine in korea. Plant Disease (ja) (2016)
37. Ng, T.F.F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., Oderinde, B.S., Wommack, K.E., Delwart, E.: High variety of known and new rna and dna viruses of diverse origins in untreated sewage. Journal of virology **86**(22), 12161–12175 (2012)
38. Perry, K.L., McLane, H., Hyder, M.Z., Dangl, G.S., Thompson, J.R., Fuchs, M.F.: Grapevine red blotch-associated virus is present in free-living vitis spp. proximal to cultivated grapevines. Phytopathology **106**(6), 663–670 (2016)
39. Clancy, S.: Rna splicing: introns, exons and spliceosome. Nature Education **1**(1), 31 (2008)

Tabela 1 Tamanhos mínimos e máximos de cada gênero

Gênero	Tamanho mínimo	Tamanho máximo
Begomovírus	2411	2959
Mastrevírus	2425	2982
Eragrovírus	2845	2854
Turncurtovírus	3044	3081
Curtovírus	3011	3180
Topocuvírus	2961	2961
Becurtovírus	2939	2960
Unclassified	2483	3308
Betassatélite	731	1552
Alphassatélite	955	1579

Tabela 2 Acrônimos dos modelos de classificação

Acrônimo do modelo de classificação	Gênero do modelo criado
GM1	Gêneros da família *Geminividae*
BM2	ORFs de begomovírus
MM3	ORFs de mastrevírus
BM4	ORFs de becurtovírus
CM5	ORFs de curtovírus
TM6	ORFs de turncurtovírus
EM7	ORFs de eragrovírus
ABM8	ORFs de alphassatélite e betassatélites

Tabela 3 Algoritmos/parâmetros que melhor se ajustaram para cada modelo proposto utilizando a validação cruzada com 10 partições. Os valores de precisão, acurácia, sensibilidade, taxa de falsos positivos e F-measure apresentados são a média ponderada das classes

Model	Algoritmo	Parâmetros Weka	Precisão	Acurácia	Sensibilidade	FP	F-measure
GM1	RF	-I 21 -K 4 -S 251	0.989	0.994	0.981	0.003	0.985
BM2	RF	-I 161 -K 5 -S 521	0.994	0.994	0.994	0.0009	0.994
MM3	RF	-I 51 -K 4 -S 351	1	1	1	0	1
BM4	RF	-I 211 -K 0 -S 141	1	1	1	0	1
CM5	RF	-I 141 -K 4 -S 381	1	1	1	0	1
TM6	MLP	-L 0.01 -M 0.5 -N 1500 -V 0 -S 0 -E 20 -H o	1	1	1	0	1
EM7	RF	-I 351 -K 7 -S 141	1	1	1	0	1
ABM8	RF	-I 31 -K 6 -S 431	1	1	1	0	1

Tabelas

Arquivos suplementares

Tabela suplementar 1

Descrição dos atributos e valores obtidos na execução do método de IG no conjunto de treinamento composto por gêneros. É apresentado o ranking criado com os atributos para que as diferenças e semelhanças entre os atributos possam ser vistas.

Tabela suplementar 2

Descrição dos atributos e valores obtidos durante a execução do método de IG em cada conjunto de treinamento composto por genes. É apresentado o ranking criado com os atributos para que as diferenças e semelhanças entre os atributos de cada conjunto de treinamento possam ser vistas.

Capítulo 5

Extração de informação de vírus de planta e importantes patógenos virais de humanos usando abordagens de processamento de linguagem natural

Artigo científico a ser submetido à revista Bioinformatics Oxford.

SOFTWARE

Extração de informação de vírus de planta e importantes patógenos virais de humanos usando abordagens de processamento de linguagem natural

Jose CF Silva[1,3†], Thales FM Carvalho[1†], Alcione Paiva[1], Elizabeth BP Fontes[2,3] and Fabio R Cerqueira[1*]

*Autor correspondente:
fabio.cerqueira@ufv.br
[1]Departamento de Informatica, Campus Universitário, Vicosa, Brazil
A lista completa de informações sobre os autores disponível no final do artigo
[†]Contribuíram igualmente

Resumo

Introdução: Vírus são agentes infecciosos que se replicam no interior das células vivas de plantas e animais. O sucesso da infecção em seres humanos por determinados vírus pode ser letal. Exemplos recentes são a epidemia de vírus ebola na África e dos vírus dengue e zika no Brasil. Em planta, os vírus causam expressivos impactos econômicos no mundo todo. Um esforço enorme por pesquisadores tem sido aplicado a fim de entender como esses patógenos evoluem ao longo do tempo e, ainda, os efeitos e os sintomas causados por esses agentes infecciosos. Todo esse esforço resultou em um enorme acervo científico que ainda é pouco explorado além da leitura individual de artigos. A técnica de processamento de linguagem natural juntamente com teoria dos grafos é uma abordagem promissora para extração de informação e conhecimento de forma automática em grandes volumes de texto. Entretanto, poucos trabalhos têm se dedicado a explorar essas técnicas e aplicá-las para extração de informação para patógenos virais, incluindo aqueles que causam sérios malefícios à saúde humana.

Resultados: A análise automática de literatura científica por meio da aplicação da técnica de processamento de linguagem natural baseado em teoria dos grafos permitiu validar por meio de análise de redes que vírus que infectam tomate são os mais estudados e que os principais sintomas apresentados por plantas são sintomas de amarelecimento e enrugamento das folhas. Também identificamos que os eventos de recombinação são os fatores entre os demais mecanismos evolutivos que mais contribuem para a alta diversidade de espécies e estruturação das populações. Além disso, identificamos que geminivírus tem sido uma alternativa para a engenharia genética e biotecnologia de plantas para a produção de vacina para seres humanos.

Conclusões: A técnica de processamento de linguagem natural baseada em grafos aplicada no presente trabalho mostrou-se uma estratégia promissora para extração de relações, recuperação de informação e descoberta de conhecimento em resumos de artigos científicos. Como resultado dos esforços aqui descritos, um módulo de busca interativo para comparação de estudos entre geminivírus e os principais patógenos virais de seres humanos foi implementado e livremente disponibilizado à comunidade científica na *Geminivirus Data Warehouse* (http://geminivirus.org:8080/geminivirusdw/search_reference.jsp).

Palavras-chaves: Natural Processing Language; Theory Graph; Geminiviruses; Ebolavirus; Graph text; PageRank; Betweenness Centrality; Recombination; Mutation; Natural selection; Drift; Migration

Introdução

Geminivírus e plantas hospedeiras são submetidos à pressão co-evolutiva contínua e esta relação resulta em expressivas perdas econômicas no mundo todo. A incidência de doenças em plantas em alguns países com economia agrícola ou emergentes em agricultura ocorre devido ao comércio ou migração humana [1]. Geminivírus são classificados pelo comitê internacional de taxonomia de vírus (*International Committee on Taxonomy of Viruses*-ICTV) em sete gêneros: *Begomovirus*, *Mastrevirus*, *Curtovirus*, *Becurtovirus*, *Eragrovirus*, *Topocuvirus* e *Turncurtovirus*. Os gêneros são classificados com base na organização genômica e inseto vetor. Os componentes genômicos consistem em um (monopartido) ou dois (bipartidos) componentes de DNA que codificam 5-7 proteínas envolvidas em replicação viral, o movimento, a transmissão e patogênese (pre-coat, reg, CP, AC5, REn, TrAp e sd). Além disso, podem associar-se a alpha ou betassatélites, que são tipos de moléculas de DNA circular que codificam apenas uma proteína, alphaC1 e betaC1, respectivamente.

Inúmeros estudos têm sido conduzidos em torno da família *Geminiviridae* devido à severidade de doenças que causam em plantas. Nos últimos anos os estudos da diversidade de espécies, distribuição geográfica, mecanismos de evolução e os mecanismos de interação desses patógenos com o hospedeiro têm aumentado expressivamente. Trabalhos recentes têm demonstrado que o gênero *Begomovirus*, em particular, é constituído de um elevado número de espécies [2]. Além disso, novas descobertas de espécies [3], relatos de vírus infectando novos hospedeiros [4] e novos gêneros vêm sendo descritos [5]. A distribuição geográfica, a estrutura da população e recombinação intra espécie também tem sido estudada [6][7]. Pesquisas dedicadas aos mecanismos da interação geminivírus-planta revelam que proteínas virais interagem com proteínas do hospedeiro e atuam em diversos processos biológicos (transcrição, replicação, movimento, transdução de sinal, metabolismo de proteínas, defesa, metabolismo, stress e transmissão) em diferentes hospedeiros [8]. O movimento do DNA viral na planta, por exemplo, é mediado por proteínas do DNA-B (NSP e MP) que atuam no transporte intracelular [9] e de célula a célula. Por outro lado, a resposta imediata do hospedeiro se dá em suprimir mecanismos de tradução global como uma estratégia de imunidade antiviral da planta [10].

Estudos de alta relevância com patógenos virais são conduzidos de maneira intensificada nas áreas da saúde e medicina no mundo todo. Em especial com aqueles que afetam a saúde humana (Dengue, Ebola, H1N1 e Zika vírus). Ao longo da história humana, surtos de epidemias ocorreram em diversos lugares do globo [11][12]. Este cenário não tem sido diferente nos dias atuais. Surtos epidêmicos ainda ocorrem no século 21, como epidemias de dengue na América central e América do sul [13], ebola na África [14], a dispersão de H1N1 no mundo [15] e, recentemente, o Zika vírus nas Américas [16]. Atualmente, milhares de estudos científicos em relação a esses patógenos foram concluídos e disponibilizados livremente em bases de dados públicas [17]. O imenso acervo de dados já existente, juntamente com os novos estudos publicados frequentemente, dificultam pesquisadores a se manterem totalmente atualizado sobre esses patógenos. Com isso, surge a necessidade de aplicação de abordagens computacionais e algoritmos para mineração e extração de informação. Assim, surgem a necessidade de aplicar técnicas computacionais avançadas como mineração de texto e processamento de linguagem natural (NLP)[18] para extrair

e relacionar de informação contidas em publicações cientificas, tendo como objetivo auxiliar na descoberta de artigos e informações que podem ser de interesse do pesquisador. A teoria dos grafos e processamento de linguagem natural têm se tornado uma estratégia importante na recuperação e extração de informação. Tradicionalmente, essas áreas do conhecimento são entendidas como áreas distintas. No entanto, pesquisas recentes sugerem que estas disciplinas podem ser ligadas para uma grande variedade de aplicações, com destaque ao processamento de grandes volumes de texto [19]. As abordagens de NLP em bioinformática ainda são pouco utilizadas. Contudo, algumas pesquisas têm se dedicado a recuperar referências usando métodos de mineração de texto e reconstruir bases de dados. Por exemplo, as interações proteína-proteína e a relação gene-doença podem ser extraídos a partir de bases de artigos científicos, como o PubMed [20].

Este trabalho dedica-se a recuperar e extrair informações relacionadas a geminivírus comparando com outros importantes patógenos virais amplamente investigados nas áreas da saúde e medicina. Para realizar esse relacionamento e descoberta de conhecimento, utilizou-se a abordagem de grafo de texto para criação de uma rede de conexões de palavras que permite mensurar quais palavras são mais representativas e/ou importantes no contexto de pesquisas de geminivírus. O grafo de texto, juntamente com as métricas betweenness centrality, degree centrality e pageRank permitem quantificar a importância de cada palavra no grafo de texto. O betweenness centrality baseia-se no número de caminhos mais curtos de todos os vértices do grafo que passam por um determinado vértice, onde um elevado betweenness pode indicar vértices que definem grupos. O Degree centrality é obtido a partir do número de arestas que incidem em um vértice. Já o PageRank é bastante utilizado para classificação de páginas web com base em algoritmo de propagação de peso [21]. Tais métricas permitem indicar quais são as palavras com maior relevância na rede e como elas relacionam entre si dentro do contexto de pesquisa em membros da família Geminiviridae. O grafo de texto permite verificar como um determinado estudo está sendo abordado, por exemplo, ao verificar os pesos e interligações de palavras como recombination, mutation, natural, selection, migration e drift que retratam os mecanismos evolutivos, é possível verificar que estudos da recombinação e mutação desses vírus no contexto de evolução são os mais estudados.

Para facilitar a utilização e interpretação do grafo e pesos, foi desenvolvida uma ferramenta web em que o pesquisador pode informar um conjunto de palavras e recebe como resposta um grafo interativo e uma tabela informando os pesos de cada palavra nos estudos de geminivírus. Essa ferramenta também permite comparar os grafos e os pesos das palavras de conjuntos de resumos de artigos sobre Dengue, Ebola, H1N1 e Zika vírus. Tal comparação pode ser analisada em dois grafos interativos, um ao lado do outro, com os pesos tabulados e um heat map (mapa de calor). As abordagens de processamento de linguagem natural baseado na teoria dos grafos permitem investigar as relações entre estudos evolutivos e caracterizam aqueles que são frequentemente estudados. Além disso, identificam a concentração desses estudos, apontam viabilidade em realizar novas pesquisas e aplicação de novas metodologias em estudos realizados com membros da família Geminiviridae. A ferramenta pode ser utilizada livremente no endereço http://www.geminivirus.org:8080/geminivirusdw/search$_reference.jsp$.

Materiais e métodos

Conjunto de dados

Resumos (*Abstracts*) de artigos científicos sobre a família *Geminiviridae* e os vírus Alpha e Betassatelites (AGV *datasets*) foram coletados na *data warehouse* geminivirus.org (1616 resumos). A fim de comparar os estudos de geminivírus com outros importantes patógenos virais, foram obtidos da base de dados PubMed 7494 resumos sobre o vírus da dengue (ADV *datasets*), 4852 sobre o vírus ebola (AEV *datasets*), 54371 sobre Influenza (AIV *datasets*), 39271 sobre Hepatite (AHV *datasets*) e 734 sobre o vírus zika (AZV datasets). Os dados são extraídos seguindo o critério de palavras chave: (i) *Geminivirus*, (ii) *Alphasatellite* e *Betasatellite*, (iii) Dengue virus, (iv) Ebola virus, (v) Influenza, (vi) Hepatitis e (vii) Zika virus.

Preprocessamento de dados

No processamento de linguagem natural, frequentemente, uma das primeiras etapas é a tokenização de sentenças do corpus que, posteriormente, podem ser agrupados em n-sequências de tokens, formando n-grams. Tantos os tokens quantos os n-grams permitem informações preliminares importantes, tais como distribuição de frequência e probabilidade condicional [18]. Neste trabalho, inicialmente foram mantidos apenas os tokens contendo caracteres alfabéticos [22]. Na sequência, o corpora passou pelo processo de tokenização, eliminação das *stopwords* (e.g., *them*, *also*, *because*), redução ao radical (stemming) por meio do algoritmo de Porter [23] e, posteriormente, geração de bigrams. Efetuou-se a radicalização visando minimizar diversos tipos de erros, reduzir o número de *bigrams* e aumentar a frequência de ocorrência (e.g, o termo *Begomoviruses* é radicalizado para *Begomovirus*). Como se sabe, os bigrams são organizados de modo que cada gram mantém o sentido e a ordem de ocorrência nas sentenças. Por exemplo, a frase "*A novel nucleocytoplasmic traffic*" daria origem aos *bigrams* $B\{$*novel, nucleocytoplasmic*$\}$ e $B\{$*nucleocytoplasmic, traffic*$\}$. A frequência com que esses *bigrams* ocorrem também foi mensurada. Para realizar esse processo, utilizou-se a biblioteca de processamento de linguagem natural NLTK *toolkit* [24].

Grafo de texto

Em processamento de linguagem natural, grafos de texto podem ser uma representação gráfica de itens de textos (documento, palavras ou sentença) e suas relações. Observa-se que esse tipo de grafo é tipicamente utilizado em pré-processamento [25], extração de relações [26] e sumarização de texto. Nós de um grafo podem representar palavras simplesmente por *tokens* ou entidades mencionadas no texto. No entanto, as arestas podem conectar *tokens* com base no sentido de leitura da sentença e ainda conectar *tokens* a uma base de conhecimento. Inicialmente, um grafo de texto G = (V,E) foi criado, onde os vértices de V(G) representam *tokens* e as arestas de E(G) representam um conjunto de *bigrams* únicos. Assim, a uma aresta e = uv representa um *bigram*. Além disto, uma aresta possui determinado pela Equação 1:

$$w = 1 - \left(\frac{\sum_{i=1}^{N} Bi}{\sum_{b=0}^{N}\sum_{i=1}^{N} Bi}\right) \tag{1}$$

onde $\sum_{i=1}^{N} Bi$ é a soma dos *bigrams* únicos e $\sum_{b=0}^{N}\sum_{i=1}^{N} Bi$ a soma total de *bigrams*. Em outras palavras, as arestas E(G) do grafo G são pontuadas pela frequência com que o *bigram* ocorre no corpus linguístico.

Análise da rede de conexões de *bigram*

A análise de redes é importante para identificar, extrair relações ou informações em grandes redes modeladas em grafos. Explorar redes de conexões para extrair conhecimento tem se tornado comum em redes sociais, biologia molecular e redes de comunicação. Essa prática é importante para conhecer a relevância de entidades objetivando identificar elementos chave (contexto de interesse). Por outro lado, para verificar elementos chave é necessário mensurar as centralidades de elementos de interesse na rede. Assim, são propostos métodos para determinar a importância com que tais elementos ocorrem na rede.

Betweenness centrality

A centralidade de intermediação no grafo de texto G = (V,E) foi calculada observando o número de vezes que um vértice v (*token*) contribui como link ao longo dos caminhos mais curtos entre todos os nós. Dessa forma, a intermediação central de um *token* v(G) é a soma da fração de todos os pares de caminhos mais curtos que passam por ele. Sendo assim, a centralidade desse vértice pode ser mensurada pela Equação 2, onde V é conjunto total de *tokens*, $\sigma(s,t)$ é o número de caminhos mais curtos de (s,t) e $\sigma(s,t|v)$ é o número de caminhos passando por v dado que (s,t) sejam *tokens* distintos [27]. Neste caso, utilizou-se o algoritmo *betweenness centrality* implementado na biblioteca *python networkx* v.1.9.1 [28].

$$\sum_{s,t\in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{2}$$

Centralidade de Graus

O grau de um vértice v representa o número de arestas incidentes. Além disso, também pode ser o número de vértices adjacentes de v. O grau máximo de um grafo G é denotado por: $\Delta(G)$ e o grau mínimo $\delta(G)$ [19]. Os graus de conexão de cada vértice v foram calculados $\sum_{i=1}^{K} di$, onde d representa cada vértice adjacente e K o número de total de vértices adjacentes.

Algoritmo PageRank

O algoritmo PageRank foi desenvolvido na Universidade de Stanford por Larry Page e Sergey Brin em 1996 [29]. Esse algoritmo analisa a estrutura do grafo G = (V,E), a fim de ranquear os nós, baseando-se nas conexões destes. Também foi originalmente projetado como um algoritmo para ranquear páginas web [30] nos mecanismos de busca do Google. Neste trabalho, o PageRank foi adotado para ranquear *tokens* que mais recebem conexões no grafo, com objetivo de identificar os *tokens* mais importantes dentro da rede. Para lograr essa finalidade, foi utilizado o algoritmo PageRank implementado na biblioteca *python networkx* v.1.9.1 [28].

Algoritmo de clustering

A análise de clustering hierárquico permite organizar subconjuntos de dados em diferentes grupos. Este método atua de forma iterativa em cada subconjunto. Em cada iteração visam-se juntar os dois agrupamentos mais semelhantes, continuando até que haja apenas um único conjunto. Em cada iteração, as distâncias entre os agrupamentos são recalculadas usando a distância euclidiana quadrada. No presente trabalho, o método utilizado para tal, foi a dissimilaridade de Lance-Williams [31] [32].

Para obter-se o agrupamento hierárquico, foram realizadas comparações par a par, em que, compararam-se os conjuntos de resumos AGV, ADV, AEV, AIV, AHV e AZV. As comparações foram realizadas agrupando *tokens* e conjuntos de dados utilizando a centralidade de graus de cada nó (*token*) do grafo. As distâncias foram padronizadas para cada variável (coluna) utilizando a métrica de normalização z-score. Tal agrupamento foi realizado com auxílio da função hclust do programa R.

Visualização gráfica

Os recursos gráficos interativos podem facilitar a visualização e interpretação das informações obtidas a partir do relacionamento de todos os resumos em uma estrutura de grafo. Uma maneira eficiente de sumarizar esses resultados é a utilização de ferramentas que possibilitam a visualização gráfica das conexões de *tokens*. Para endereçar essa restrição utilizou-se a ferramenta cytoscape.js para geração gráfica e interativa de um grafo. A visualização gráfica do agrupamento dos conjuntos de dados foram obtidos usando a função do R heatmap.2 e a biblioteca gráfica ggplot2 [33]. Em síntese, o processo de análise inicia-se adquirindo as palavras de um resumo de geminivírus. Na sequência obteve-se o valor da centralidade de graus do grafo, caso existam no grafo. Esse processo também se repete para os demais conjuntos de dados. Posteriormente, uma matriz com valores de centralidade de graus (palavras vs conjunto de dados) é obtida. Então, os métodos de agrupamento são aplicados e os resultados são visualizados em formato de mapa de calor.

Resultado e discussão

Grafo de texto

A coleção de *bigrams* obtidos foi utilizada para construir redes de conexões de palavras, para cada conjunto de dados. Os *bigrams* mais frequentes no *corpus* de geminivírus estão relacionados com sintomas de plantas, hospedeiros, classificação viral, material genético (DNA) do vírus e localização geográfica (Tabela 1).

É possível observar na Tabela 1 que, em sua maioria, pesquisas que envolvem a família *Geminiviridae* concentram-se em classificação taxonômica, sintomas de doenças causadas em plantas, o vetor de infecção viral e análise de sequência genômica. Ao analisar os *bigrams*, é facilmente notado que o *bigram* $B\{leaf,curl\}$ obteve a maior ocorrência entre todos os *bigrams*, representando duas vezes mais o segundo colocado $B\{mosaic,virus\}$, sugerindo que esses são os principais sintomas de plantas quando infectadas por geminivirus. Além disso, é possível notar que entre os 20 *bigrams* melhores ranqueados, o *token* representado pela palavra *tomato* se repete em dois diferentes pares de *bigrams*, $B\{tomato,yellow\}$ e $B\{tomato,leaf\}$.

Estes representam 11,55% do total dos *bigrams* melhores ranqueadas. Também os *bigrams* B{*cassava,mosaic*} e B{*cotton,leaf*} estão entre os melhores ranqueados. É importante dar atenção a tais ocorrências, devido ao fato de o tomate, a mandioca e o algodão estarem entre as principais commodities mais cultivadas em diferentes continentes nas últimas décadas, segundo dados apurados pelo *Food And Agriculture Organization Of The United Nations Statistics Division* (faostat3.fao.org). Em adição, doenças causadas por geminivírus nestes cultivares causaram sérios impactos econômicos e sociais na Europa e Ásia (*tomato e cotton leaf curl diseases*, TYLCD e CLCuD; respectivamente) [34][35] e Africa (*cassava mosaic disease* e CMD) [36].

Análise da rede

As métricas de análise de redes são fundamentais para identificar entidades importantes no grafo. Métricas de centralidade em grafos como PageRank, *betweenness centrality* e centralidade de graus são utilizados para identificar a influência de pessoas dentro de uma rede social [37] e/ou rankear as páginas da web, e.g., o PageRank mensura a importância de uma página web com base nos links de entrada de outras páginas que permitem acesso a mesma[30]. Essas métricas aplicadas a um grafo de texto permitem identificar quais *tokens* são mais importantes dentro de um conjunto de resumos de artigos. Para avaliar o potencial do grafo e extrair informações, selecionou-se um conjunto de cinco palavras que destacam os mecanismos evolutivos na família *Geminiviridae* (Figura 1). Os mecanismos são avaliados por meio de *tokens* de maneira contextual, em que cada *token* representa um contexto de estudos já realizados sobre essa família. O *token recombination* contextualiza pesquisas realizadas que envolveram a troca aleatória de material genético. O *token mutation* refere-se a mudanças na sequência dos nucleotídeos do material genético de um organismo. O *token migration* representa a dinâmica com que indivíduos migraram geograficamente incorporando-se ou retirando-se de uma população. A seleção natural é representada pelos *tokens natural* e *selection* os quais indicam a adaptação e a especialização dos seres vivos [38]. Por fim, o mecanismo deriva genética é representado pelo *token drift*, que direciona a ideia do mecanismo microevolutivo que modifica aleatoriamente as frequências alélicas ao longo do tempo.

Por meio da análise da rede é possível mensurar a importância desses cinco mecanismos evolutivos no conjunto AGV a partir das métricas apresentadas na Tabela 2. Entre os cinco mecanismos selecionados, o *token recombination* obteve os maiores valores no ranking nas três métricas aplicadas. Isso sugere que a recombinação é o mecanismo de evolução mais estudados na família *Geminiviridae* [39][40]. Além disso, a importância desse *token* demonstra que a recombinação é provavelmente o tópico que mais tem sido estudado por geminivirologistas em relação aos demais mecanismos. Ainda utilizando as métricas apresentadas da Tabela 2, é possível observar que a mutação é o segundo mecanismo evolutivo que contribui para a rápida evolução e alta diversidade de espécies [41]. Os *tokens* natural e *selection* também são bem ranqueados em relação aos *tokens migration* e *drift*. Portanto, isso nos leva a entender que esses estudos também são frequentemente estudados. Entretanto, avaliando a frequência com que essas *tokens* ocorrem simultaneamente (*bigram* B{*natural,selection*}), a frequência é relativamente baixa em relação às ocorrências destes *tokens* separadamente. Tal fato revela que esses últimos mecanismos ainda são poucos explorados em pesquisas com geminivírus.

Uma maneira interativa de analisar os *tokens* é navegar intuitivamente pelo grafo. Assim, é possível analisar as pontuações obtidas pelas métricas de centralidade de cada nó. Além disso, o número de ocorrência e a frequência de cada *bigram* podem ser visualizados. Ao avaliar os nós vizinhos de cada um desses mecanismos anteriormente analisados, percebe-se que os nós *recombination* e *mutation* possuem o maior número de vizinhos, seguidos de seleção (*selection*), migração (*migration*) e deriva (*drift*). Essa maneira de analisar também possibilita alcançar os mesmos os resultados da Tabela 2. O acesso à navegação interativa pelo grafo é livremente acessível através do *Geminivirus Data warehouse* na sessão *publication* (www.geminivirus.org:8080/geminivirusdw/search_reference.jsp).

Ao realizar a análise da rede sobre o conjunto AGV utilizando palavras relacionadas a mecanismos evolutivos, percebeu-se grande potencial em utilizar a modelagem de grafo de texto para comparar pesquisas de diferentes áreas. Sendo assim, aplicou-se todo o processo de criação do grafo para geminivírus (filtragem, tokenização, visualização, etc) em outros cinco conjuntos de *abstract* de patógenos amplamente estudados (ADV, AEV, AIV, AHV e AZV). Esses grafos de texto possibilitam realizar uma análise comparativa entre essas áreas de estudos de interesse. Entende-se que a realização dessas comparações dependerá da demanda de cada pesquisador e do contexto de busca ou pergunta que se deseja responder baseado na recuperação da informação ou conhecimento. Por exemplo, para identificar relações entre estudos realizados simultaneamente entre vírus ebola e geminivírus, ou contextos que são comuns entre essas duas áreas, realizaram-se buscas por esses contextos simultaneamente nos dois grafos. Assim, os escores obtidos nas métricas de centralidade permitiram identificar palavras chave importantes nas redes. Com o objetivo de extrair informações que relacionassem essas duas áreas de estudos, algumas palavras chaves dessas áreas foram submetidas a tal comparação. Inicialmente, utilizaram-se os mesmos *tokens* (*recombination*, *mutation*, *natural*, *selection*, *migration* e *drift*) que relatam a ocorrência de palavras que representam os mecanismos evolutivos citados anteriormente. Também foram acrescentados *tokens* que pudessem revelar a ocorrência nos nós grafos (ebola, *Geminivirus* e *geminiviral*) (Figura 2). Comparando os dois grafos é possível inferir, baseando-se nos resumos publicados, que eventos de recombinação e mutação ocorrem mais frequentemente em geminivírus. Da mesma forma, as palavras seleção e natural ocorrem com maior frequência, entretanto, quando analisadas em modelo *bigram*, forma-se o contexto de seleção natural, o qual ocorre com baixa frequência. Também é possível inferir que a migração ocorre mais vezes no grafo de vírus ebola e suas medidas de centralidade possuem maior escore.

Avaliando as medidas de centralidade das palavras ebola, *Geminivirus* e *geminiviral*, verificou-se que tais *tokens* ocorrem simultaneamente nos dois grafos, uma vez que esses patógenos possuem hospedeiros e mecanismos de infecção completamente diferentes. Os nós (*tokens*) *Geminivirus* e *geminiviral* do grafo AGV possuem maior escore em relação ao grafo de AEV. Onde, os escores de centralidade de graus do grafo AGV possui valores 1382 e 146 e o grafo de AEV 6 e 2 respectivamente. Entretanto, a palavra ebola ocorre também com baixa frequência no grafo de AGV, tendo escore PageRank com valor 0.000073214, *betweenness centrality* com valor 0.0002826683 e centralidade de graus com valor 15. Sugere-se que há uma

relação entre esses dois patógenos quando comparados por meio desses metadados e a aplicação de técnicas de mineração de dados. Os mecanismos de busca em resumos de artigos na base de dados *Geminivirus Data Warehouse* (geminivirus.org) foram adaptados para receber requisições de busca oriundas da navegação interativa do grafo. Esse mecanismo permitiu submeter buscas contextuais dentro dessa base de dados e selecionar os resumos que representam esses contextos. Neste sentido, navegando intuitivamente pelo grafo AGV, submeteu-se o contexto ebola, a fim de realizar uma busca desse contexto nas bases de dados de resumos específicos de geminivírus. O resultado dessa busca ranqueou cinco importantes resumos de artigos que demonstram que geminivírus tem sido frequentemente utilizado em estratégias da engenharia genética e biotecnologias de plantas para a produção de vacinas para seres humanos [42][43][44][45][46]. Em adição, nestes trabalhos utilizou-se um sistema de replicação *geminiviral* para produzir um complexo imune para ebola em plantas de *Nicotiana benthamiana* [42]. Também demonstram que a alta eficiência que geminivírus, em particular *bean yellow dwarf virus* (BeYDV), tem em replicar proteínas para produção de vacinas [43]. Além disso, demonstra potencial para produção de anticorpo contra vírus ebola [44]. Abordagens similares também têm sido utilizadas para produção de vacinas contra dengue [45], demostrando grande potencial da utilização de vetor de replicação baseado em geminivírus com aplicação em biotecnologia de plantas e expressando potencial da sua utilização como agente farmacêutico biológico.

Outra maneira interessante de representar graficamente palavras é a utilização de mapa de calor com palavras agrupadas por meio de cluster hierárquico usando a centralidade de grau. Assim, o resultado da busca por palavra chave ou usando todas as palavras do resumo é avaliado graficamente através da intensidade em que estas palavras aparecem em todos os conjuntos de dados de resumos (Figura 3). A interpretação do mapa de calor se dá em relação aos demais conjuntos de dados. Os algoritmos de agrupamentos agrupam as palavras que ocorrem de forma similar. Além disso, os conjuntos de dados também são associados, sugerindo que as palavras do resumo ocorrem com frequência similar.

Conclusão

Os resultados dessa pesquisa sugerem que a técnica de processamento de linguagem natural baseado em teoria dos grafos pode ser uma estratégia promissora para extração de relações e conhecimento em resumos de artigos científicos. Também sugerem que a centralidade de interligações, o ranqueamento de nós pelo algoritmo PageRank e também a análise de centralidade de graus podem determinar a relevância com que as palavras ocorrem no corpus. Ao fazer o relacionamento do resultado dessas métricas, pode-se determinar o grau de relevância e quais palavras ou entidades são mais relevantes nestes textos.

A solução proposta neste trabalho permite extrair relações e informações baseando-se não apenas na contagem de palavras e frequência em que as mesmas ocorrem, mas também na importância e relevância detectada nos textos científicos que descrevem pesquisas já consolidadas. O *betweenness centrality* permite o entendimento do quanto uma palavra pode contribuir para conectar vias de menor tamanho na rede. O ranqueamento de palavras pelo algoritmo PageRank contribuiu

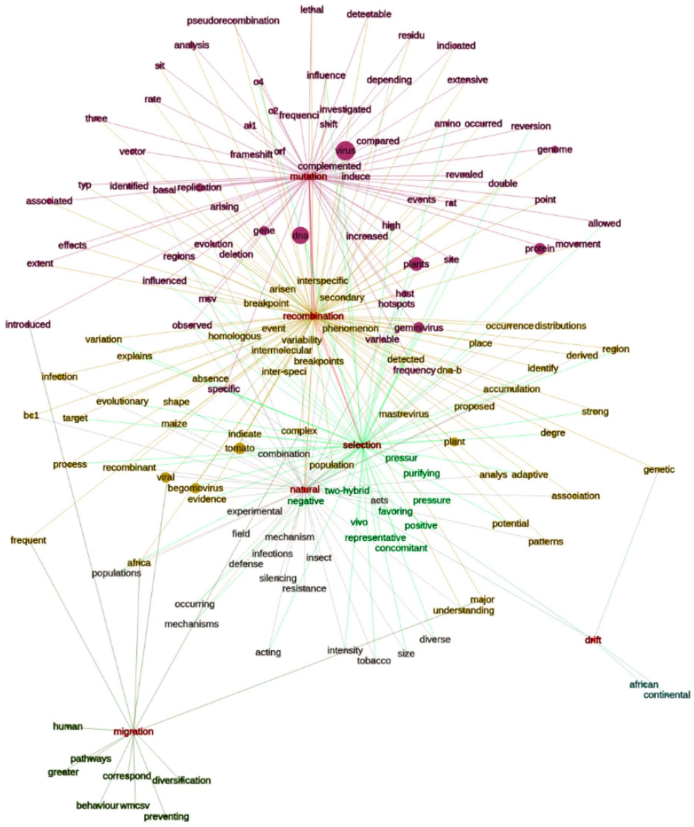

Figura 1 Grafo de palavras chave que representam os mecanismos evolutivos Apresenta um grafo de palavras é formado por meio de palavras chave. As palavras chave estão mascadas em vermelho. Cada palavra chave está conectada a nós vizinhos por meio de arestas com cores diferentes.

no esclarecimento de quais palavras seriam mais importantes dentro do grafo. Cada palavra direciona o seu sentido gramatical, permitindo que os contextos de sua ocorrência dentro de frases garantam o sentido semântico. Essa metodologia também pode proporcionar de forma resumida a revisão de literatura rápida e permitir que pesquisadores conheçam quais áreas de estudos são mais frequentemente estudadas e mais exploradas ou identificar novas oportunidades de aplicar técnicas ou, ainda, desenvolvimento de novos estudos. Pesquisadores interessados em extrair informações sobre geminivírus, e comparar contextos com outros importantes patógenos virais mencionados, podem acessar livremente os mecanismos de buscas implementados na *Geminivirus Data Warehouse* (http://www.geminivirus.org:8080/geminivirusdw/search_reference.jsp).

Abreviações
NLP: natural language processing **AGV:** datasets of *Geminivirus* abstracts **AEV:** datasets of Ebolavirus abstracts **AIV:** datasets of Influenza abstracts **AHV:** datasets of Hepatitis abstracts **AZV:** datasets of Zica virus abstracts **ADV:** datasets of Dengue virus abstracts

Agradecimentos
Os autores agradecem a Universidade Federal de Viçosa (UFV), Instituto Nacional de Ciência e Tecnologia em Interações Planta-Praga (INCTIPP), Departamento de Informática - UFV e Divisão de Suporte ao Desenvolvimento Científico e Tecnológico - UFV. Agradecimentos pelo apoio financeiro: CAPES, CNPq, FAPEMIG and INCTIPP.

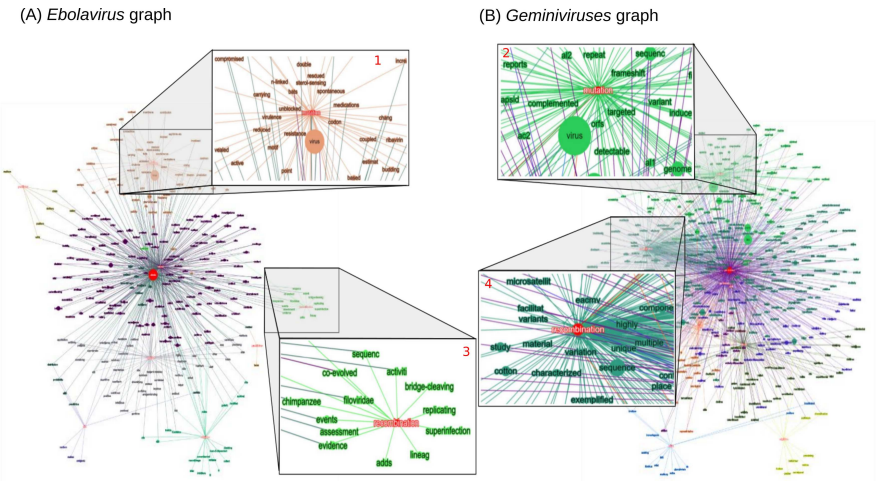

Figura 2 : Ilustração da comparação entre grafos de geminivírus e vírus ebola A figura apresenta dois grafos: o primeiro grafo (A) gerado pelo conjunto de dados de vírus ebola e o segundo grafo (B) gerado pelo conjunto de dados de geminivírus. Os quadros 1 e 2 apresentam as conexões do nó/token que representam mutação. Os nós 3 e 4 mostram as conexões dos nós que representam o contexto recombinação.

Detalhes dos autores

[1]Departamento de Informatica, Campus Universitário, Vicosa, Brazil. [2]Departamento de Bioquimica, Campus Universitário, , Vicosa, Brazil. [3]Instituto Nacional de Ciência e Tecnologia em Interações Planta-Praga/BIOAGRO, Campus Universitário, , Vicosa, Brazil.

Referências

1. Khan, A.J., Mansoor, S., Briddon, R.W.: Oman: a case for a sink of begomoviruses of geographically diverse origins. Trends in plant science **19**(2), 67–70 (2014)
2. Brown, J.K., Zerbini, F.M., Navas-Castillo, J., Moriones, E., Ramos-Sobrinho, R., Silva, J.C., Fiallo-Olivé, E., Briddon, R.W., Hernández-Zepeda, C., Idris, A., *et al.*: Revision of begomovirus taxonomy based on pairwise sequence comparisons. Archives of virology **160**(6), 1593–1619 (2015)
3. Nascimento, L.D., Silva, S.J., Sobrinho, R.R., Ferro, M.M., Oliveira, M.H., Zerbini, F.M., Assunção, I.P., Lima, G.S.: Complete nucleotide sequence of a new begomovirus infecting a malvaceous weed in brazil. Archives of virology **161**(6), 1735–1738 (2016)
4. Leke WN, B.J.F.V. Mignouna DB: First report of soybean chlorotic blotch virus and west african asystasia virus 1 infecting cassava and a wild cassava relative in cameroon and togo. New Disease Reports **33**(24) (2016)
5. Roumagnac, P., Granier, M., Bernardo, P., Deshoux, M., Ferdinand, R., Galzi, S., Fernandez, E., Julian, C., Abt, I., Filloux, D., *et al.*: Alfalfa leaf curl virus: An aphid-transmitted geminivirus. Journal of virology **89**(18), 9683–9688 (2015)
6. Rocha, C.S., Castillo-Urquiza, G.P., Lima, A.T., Silva, F.N., Xavier, C.A., Hora-Júnior, B.T., Beserra-Júnior, J.E., Malta, A.W., Martin, D.P., Varsani, A., *et al.*: Brazilian begomovirus populations are highly recombinant, rapidly evolving, and segregated based on geographical location. Journal of virology **87**(10), 5784–5799 (2013)
7. Lima, A., Ramos-Sobrinho, R., Gonzalez-Aguilera, J., Rocha, C., Silva, S., Xavier, C., Silva, F., Duffy, S., Zerbini, F.: Synonymous site variation due to recombination explains higher variability in begomovirus populations infecting non-cultivated hosts. Journal of General Virology **94**, 418–431 (2012)
8. Hanley-Bowdoin, L., Bejarano, E.R., Robertson, D., Mansoor, S.: Geminiviruses: masters at redirecting and reprogramming plant processes. Nature Reviews Microbiology **11**(11), 777–788 (2013)
9. Florentino, L.H., Santos, A.A., Fontenelle, M.R., Pinheiro, G.L., Zerbini, F.M., Baracat-Pereira, M.C., Fontes, E.P.: A perk-like receptor kinase interacts with the geminivirus nuclear shuttle protein and potentiates viral infection. Journal of virology **80**(13), 6648–6656 (2006)
10. Zorzatto, C., Machado, J.P.B., Lopes, K.V., Nascimento, K.J., Pereira, W.A., Brustolini, O.J., Reis, P.A., Calil, I.P., Deguchi, M., Sachetto-Martins, G., *et al.*: Nik1-mediated translation suppression functions as a plant antiviral immunity mechanism. Nature **520**(7549), 679–682 (2015)
11. Serfling, R.E.: Historical review of epidemic theory. Human biology **24**(3), 145–166 (1952)
12. Uribe, J.: Contagion: Historical views of diseases and epidemics. Nursing History Review **18**, 207 (2010)
13. Guzman, G., Kouri, G., *et al.*: Dengue and dengue hemorrhagic fever in the americas: lessons and challenges. Journal of Clinical Virology **27**(1), 1–13 (2003)
14. Goeijenbier, M., Van Kampen, J., Reusken, C., Koopmans, M., Van Gorp, E.: Ebola virus disease: a review on epidemiology, symptoms, treatment and pathogenesis. Neth J Med **72**(9), 442–8 (2014)
15. Dawood, F.S., Iuliano, A.D., Reed, C., Meltzer, M.I., Shay, D.K., Cheng, P.-Y., Bandaranayake, D., Breiman, R.F., Brooks, W.A., Buchy, P., *et al.*: Estimated global mortality associated with the first 12 months of 2009

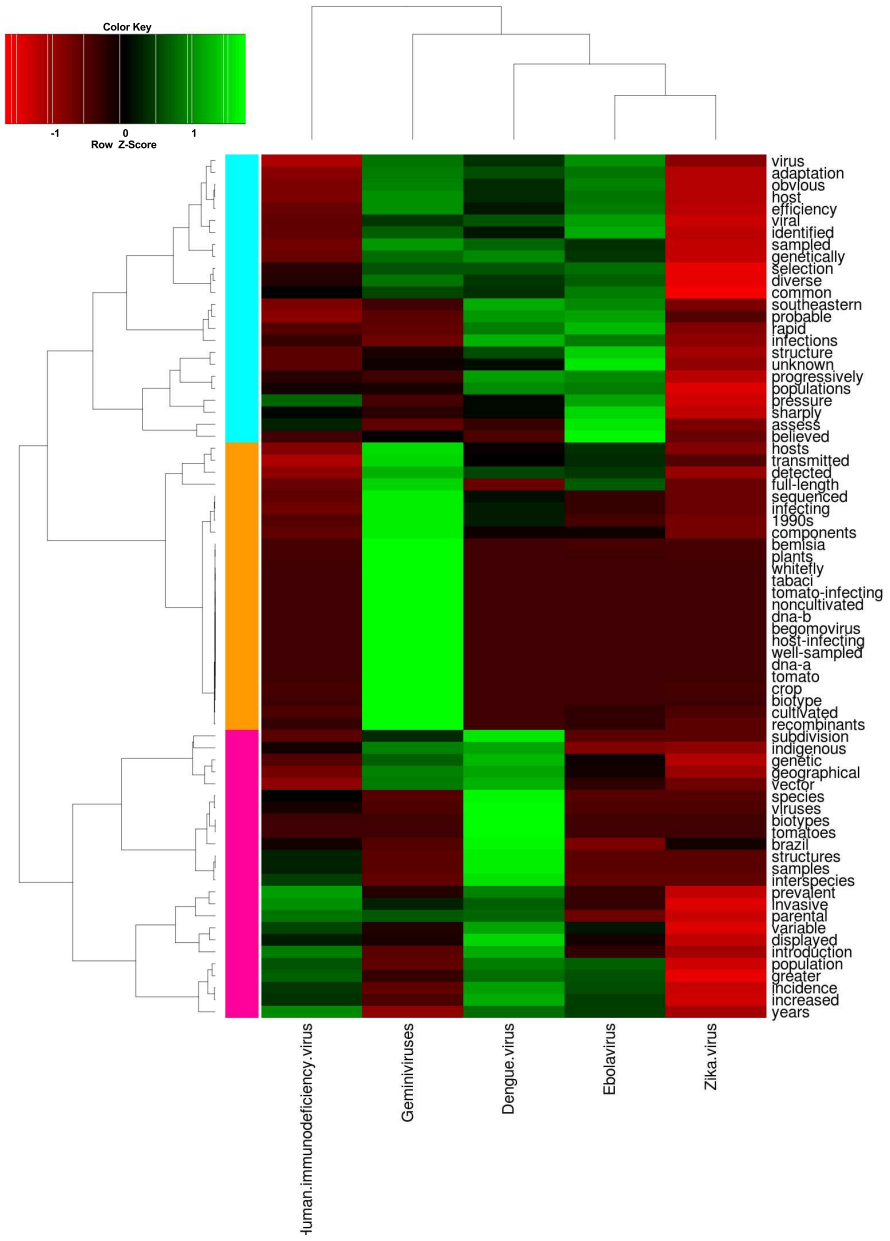

Figura 3 Mapa de calor para a clusterização de todas as palavras de um resumo O agrupamento apresenta grupos mais próximos de palavras coloridos em azul, laranja e rosa. As cores verdes apontam as palavras com maior grau no grafo, enquanto que as cores vermelhas mostram aquelas que menos ocorrem. As demais cores apontam aquelas com valores de centralidade de graus com valores intermediários.

pandemic influenza a h1n1 virus circulation: a modelling study. The Lancet infectious diseases **12**(9), 687–695 (2012)

16. Petersen, E., Wilson, M.E., Touch, S., McCloskey, B., Mwaba, P., Bates, M., Dar, O., Mattes, F., Kidd, M., Ippolito, G., *et al.*: Rapid spread of zika virus in the americas-implications for public health preparedness for mass gatherings at the 2016 brazil olympic games. International Journal of Infectious Diseases **44**, 11–15 (2016)
17. ncbi: Pubmed. pubmed.com **2015** (2015)
18. Jurafsky, D., James, H.: Speech and language processing an introduction to natural language processing,

computational linguistics, and speech (2000)
19. Mihalcea, R., Radev, D.: Graph-based Natural Language Processing and Information Retrieval. Cambridge University Press, New York (2011)
20. Zeng, Z., Shi, H., Wu, Y., Hong, Z.: Survey of natural language processing techniques in bioinformatics. Computational and mathematical methods in medicine **2015**, 1 (2015)
21. Yan, E., Ding, Y.: Applying centrality measures to impact analysis: A coauthorship network analysis. Journal of the American Society for Information Science and Technology **60**(10), 2107–2118 (2009)
22. Hieronymus, J.L.: Ascii phonetic symbols for the world's languages: Worldbet. Journal of the International Phonetic Association **23** (1993)
23. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980)
24. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. " O'Reilly Media, Inc.", Cambridge Mass (2009)
25. Reimer, U., Hahn, U.: Text condensation as knowledge base abstraction. In: Artificial Intelligence Applications, 1988., Proceedings of the Fourth Conference On, pp. 338–344 (1988). IEEE
26. Melli, G.S.: Supervised ontology to document interlinking. PhD thesis, Applied Science: School of Computing Science (2010)
27. Brandes, U.: On variants of shortest-path betweenness centrality and their generic computation. Social Networks **30**(2), 136–145 (2008)
28. Schult, D.A., Swart, P.: Exploring network structure, dynamics, and function using networkx. In: Proceedings of the 7th Python in Science Conferences (SciPy 2008), vol. 2008, pp. 11–16 (2008)
29. Vice, D., Malseed, M.: The Google Story. Delacorte Press [Random House] New York, NY: (2005)
30. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: bringing order to the web. (1999)
31. Lance, G.N., Williams, W.T.: A general theory of classificatory sorting strategies ii. clustering systems. The computer journal **10**(3), 271–277 (1967)
32. Murtagh, F., Legendre, P.: Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? Journal of Classification **31**(3), 274–295 (2014)
33. Wickham, H.: Ggplot2: Elegant Graphics for Data Analysis. Springer, New York (2009). http://ggplot2.org
34. Nannini, M., Foddi, F., Murgia, G., Pisci, R., Sanna, F., Testa, M., Accotto, G.: An epidemiological survey of tylcd in southern sardinia (italy). Communications in agricultural and applied biological sciences **74**(3), 831–841 (2008)
35. Sanz, A.I., Fraile, A., Gallego, J.M., Malpica, J.M., García-Arenal, F.: Genetic variability of natural populations of cotton leaf curl geminivirus, a single-stranded dna virus. Journal of molecular evolution **49**(5), 672–681 (1999)
36. Alabi, O.J., Ogbe, F.O., Bandyopadhyay, R., Kumar, P.L., Dixon, A.G., Hughes, J.d., Naidu, R.A.: Alternate hosts of african cassava mosaic virus and east african cassava mosaic cameroon virus in nigeria. Archives of virology **153**(9), 1743–1747 (2008)
37. Otte, E., Rousseau, R.: Social network analysis: a powerful strategy, also for the information sciences. Journal of information Science **28**(6), 441–453 (2002)
38. Charles, D.: The Origin of Species. John Murry London, London (1929)
39. Padidam, M., Sawyer, S., Fauquet, C.M.: Possible emergence of new geminiviruses by frequent recombination. Virology **265**(2), 218–225 (1999)
40. Lima, A.T., Sobrinho, R.R., Gonzalez-Aguilera, J., Rocha, C.S., Silva, S.J., Xavier, C.A., Silva, F.N., Duffy, S., Zerbini, F.M.: Synonymous site variation due to recombination explains higher genetic variability in begomovirus populations infecting non-cultivated hosts. Journal of General Virology **94**(2), 418–431 (2013)
41. Arguello-Astorga, G., Ascencio-Ibáñez, J.T., Dallas, M.B., Orozco, B.M., Hanley-Bowdoin, L.: High-frequency reversion of geminivirus replication protein mutants during infection. Journal of virology **81**(20), 11005–11015 (2007)
42. Phoolcharoen, W., Bhoo, S.H., Lai, H., Ma, J., Arntzen, C.J., Chen, Q., Mason, H.S.: Expression of an immunogenic ebola immune complex in nicotiana benthamiana. Plant biotechnology journal **9**(7), 807–816 (2011)
43. Chen, Q., He, J., Phoolcharoen, W., Mason, H.S.: Geminiviral vectors based on bean yellow dwarf virus for production of vaccine antigens and monoclonal antibodies in plants. Human vaccines **7**(3), 331–338 (2011)
44. Huang, Z., Phoolcharoen, W., Lai, H., Piensook, K., Cardineau, G., Zeitlin, L., Whaley, K.J., Arntzen, C.J., Mason, H.S., Chen, Q.: High-level rapid production of full-size monoclonal antibodies in plants by a single-vector dna replicon system. Biotechnology and bioengineering **106**(1), 9–17 (2010)
45. Kim, M.-Y., Reljic, R., Kilbourne, J., Ceballos-Olvera, I., Yang, M.-S., Reyes-del Valle, J., Mason, H.S.: Novel vaccination approach for dengue infection based on recombinant immune complex universal platform. Vaccine **33**(15), 1830–1838 (2015)
46. Lai, H., He, J., Engle, M., Diamond, M.S., Chen, Q.: Robust production of virus-like particles and monoclonal antibodies with geminiviral replicon vectors in lettuce. Plant biotechnology journal **10**(1), 95–104 (2012)

Tabelas

Tabela 1 Frequência de *Bigrans*

Token	Token	Frequência
leaf	curl	1654
mosaic	virus	728
yellow	leaf	627
tomato	yellow	622
curl	virus	576
viral	dna	390
cassava	mosaic	364
tomato	leaf	344
nicotiana	benthamiana	334
coat	protein	319
golden	mosaic	266
yellow	mosaic	265
dna	replication	245
bemisia	tabaci	244
yellow	vein	244
cotton	leaf	236
nucleotide	sequence	236
sequence	identity	226
family	geminiviridae	224
gene	silencing	214

Tabela 2 Ranking dos mecanismos evolutivos na rede

Mecanismo	Centralidade	PageRank	Grau height
recombination	0,0115124	0,002304	498
mutation	0,002647	0,0006474	188
natural	0,0010756	0,0005466	135
selection	0,0014598	0,0003758	114
migration	1,955E-05	3,901E-05	14
drift	1,597E-06	2,126E-05	5

Capítulo 6

Conclusões

A identificação de gêneros da família *Geminiviridae* tem se tornado um desafio devido ao elevado número de sequências disponíveis em bases de dados. Além disso, um grande conhecimento em taxonomia e análises de bioinformática é atualmente requerido. Para solucionar essas restrições, apresentou-se um Data Warehouse designado geminivirus.org. Um Data Warehouse enriquecido com abordagens de aprendizado de máquina, de acesso aberto, com dados integrados e que otimiza pesquisas abrangentes que são difíceis de executar usando ferramentas existentes. Sua flexibilidade permite a análise de vários tipos de informações sobre geminivírus tais como taxonomia, sequência genômica e recuperação de dados biológicos. Os mecanismos de busca e ferramentas analíticas facilitam pesquisas complexas para pesquisadores que não possuem conhecimento em análise computacional de dados. Além disso, auxiliam pesquisadores em suas tomadas de descisão de maneira simples e amígavel. O fato de ter sido projetado utilizando padrões e conceitos de reusabilidade de códigos permite que novas funcionalidades e metodologias sejam facilmente incorporadas. Além de ser uma base de dados que flexibiliza a adição de sequências genômicas de novas famílias ou novos gêneros recém descobertos.

Como um dos principais resultados dessa pesquisa, propõe-se um poderoso método de predicão e classificação de genes para Geminivírus designado Fangorn Forest (F2). Como parte desse método, o algoritmo Millau Bridge (MB) também foi proposto, designado assim porque as estratégias utilizadas foram inspiradas na construção de uma importante ponte francesa que possui o mesmo nome. O

algoritmo MB utiliza uma abordagem gulosa para identificar as ORFs baseando-se nos codons de iniciação e codons de terminação dos cDNAs. Além do mais, prediz sítios de splicing visando encontrar sítios de ramificação (*branch point*), sítios GU e AG (regra GU-AG) e as regiões ricas em pirimidinas na extremidade do intron, sendo essas as principais características para lograr o sucesso do algoritmo em relação às ferramentas existentes. Como pode ser visto, os modelos de ML para classificação desses genes superam 99% de acurácia, precisão e sensitividade. A alta precisão foi alcançada devido a utlização de características extraídas das sequências de CDS e sequências de aminoácidos. Tais características são comuns a qualquer sequencia de CDS ou aminoácidos. Isso reforça a idéia de que o método pode ser facilmente expandido para outros vírus de DNA circular ou bactérias.

Os resumos de artigos também foram analisados aplicando-se poderosas técnicas de NLP. Algumas hipóteses foram geradas utilizando essa técnica, que deu origem a um dos principais trabalhos de pesquisa aqui concretizados,como, por exemplo, a aplicação de aprendizado de máquina para classificar genes da família *Geminiviridae*. Entretanto, outras perguntas foram respondidas utilizando essa abordagem. Ao visitar o grafo de texto por meio das palavras chave que representam os cinco mecanismos evolutivos, foi possível inferir que eventos de recombinação são os tipos de análise de dados e o mecanismo evolutivo mais estudados em geminivírus. Pode-se notar que, o potencial da utilização do método de NLP se estende como uma maneira de gerar novas hipóteses ou identificar eventuais oportunidades de novas pesquisas a serem realizadas e que ainda não foram vislumbradas.

O DW geminivirus.org é uma ferramenta que pode ser utilizada por pesquisadores em carater estratégico em suas atividades de pesquisa. Assim, o acesso completo, fácil, rápido e confiável às informações consolidam-se como uma marca crucial dessa DW. Isto, porque está sustentado pela utilização de algorítmos precisos, robustas técnicas de aprendizado de máquina e metodologias inovadoras na maneira com que os dados biológicos dessa família são trabalhados.

Referências Bibliográficas

Alberts, B.; Bray, D.; Lewis, J.; Raff, M.; Roberts, K.; Watson, J. D. & Grimstone, A. (1995). Molecular biology of the cell (3rd edn). *Trends in Biochemical Sciences*, 20(5):210--210.

Basso, M. F.; da Silva, J. C. F.; Fajardo, T. V. M.; Fontes, E. P. B. & Zerbini, F. M. (2015). A novel, highly divergent ssdna virus identified in brazil infecting apple, pear and grapevine. *Virus research*, 210:27--33.

Berry, M. J. & Linoff, G. (1997). *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, Inc.

Bird, S.; Klein, E. & Loper, E. (2009). *Natural language processing with Python*. "O'Reilly Media, Inc.".

Boehnlein, M. & Ulbrich-vom Ende, A. (1999). Deriving initial data warehouse structures from the conceptual data models of the underlying operational information systems. Em *Proceedings of the 2nd ACM international workshop on Data warehousing and OLAP*, pp. 15--21. ACM.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5--32.

Brown, J. K.; Zerbini, F. M.; Navas-Castillo, J.; Moriones, E.; Ramos-Sobrinho, R.; Silva, J. C.; Fiallo-Olivé, E.; Briddon, R. W.; Hernández-Zepeda, C.; Idris, A. et al. (2015). Revision of begomovirus taxonomy based on pairwise sequence comparisons. *Archives of virology*, 160(6):1593--1619.

Daniel, J. & James, H. (2009). Speech and language processing: An introduction to natural language processing. *Computational Linguistics and Speech Recognition, 2nd Ed., Prentice Hall*.

Frank, E.; Hall, M.; Trigg, L.; Holmes, G. & Witten, I. H. (2004). Data mining in bioinformatics using weka. *Bioinformatics*, 20(15):2479--2481.

Gardner, M. W. & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)âa review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14):2627--2636.

Inoue-Nagata, A. K.; Albuquerque, L. C.; Rocha, W. B. & Nagata, T. (2004). A simple method for cloning the complete begomovirus genome using the bacteriophage φ29 dna polymerase. *Journal of virological methods*, 116(2):209--211.

Lu, J.; Hales, A.; Rew, D.; Keech, M.; Fröhlingsdorf, C.; Mills-Mullett, A. & Wette, C. (2015). Data mining techniques in health informatics: a case study from breast cancer research. Em *International Conference on Information Technology in Bio-and Medical Informatics*, pp. 56--70. Springer.

Maimon, O. & Rokach, L. (2005). *Data mining and knowledge discovery handbook*, volume 2. Springer.

Martin, J. H. & Jurafsky, D. (2000). Speech and language processing. *International Edition*, 710.

Mena, J. (1999). *Data mining your website*. Digital Press.

Mihalcea, R. & Radev, D. (2011). *Graph-based natural language processing and information retrieval*. Cambridge University Press.

Mitchell, A.; Chang, H.-Y.; Daugherty, L.; Fraser, M.; Hunter, S.; Lopez, R.; McAnulla, C.; McMenamin, C.; Nuka, G.; Pesseat, S. et al. (2015). The interpro protein families database: the classification resource after 15 years. *Nucleic acids research*, 43(D1):D213--D221.

Platt, J. C. (1999). 12 fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods*, pp. 185--208.

Prasanna, H. & Rai, M. (2007). Detection and frequency of recombination in tomato-infecting begomoviruses of south and southeast asia. *Virology Journal*, 4(1):1.

Purusothaman, G. & Krishnakumari, P. (2015). A survey of data mining techniques on risk prediction: heart disease. *Indian Journal of Science and Technology*, 8(12):1.

Rocha, C. S.; Santos, A. A.; Machado, J. P. B. & Fontes, E. P. (2008). The ribosomal protein l10/qm-like protein is a component of the nik-mediated antiviral signaling. *Virology*, 380(2):165--169.

Souza, A. P. & Zaia, J. E. (2015). O uso do data mining na promoção de saúde: uma revisão sistemática da literatura. *Atas de Saúde Ambiental-ASA*, 3(1):12--21.

Varsani, A.; Navas-Castillo, J.; Moriones, E.; Hernández-Zepeda, C.; Idris, A.; Brown, J. K.; Zerbini, F. M. & Martin, D. P. (2014). Establishment of three new genera in the family geminiviridae: Becurtovirus, eragrovirus and turncurtovirus. *Archives of virology*, 159(8):2193--2203.

Wang, J. T.; Zaki, M. J.; Toivonen, H. T. & Shasha, D. (2005). Introduction to data mining in bioinformatics. Em *Data Mining in Bioinformatics*, pp. 3--8. Springer.

Weis, S. M. & Indurkhya, N. (1999). Predict data mining.

Westphal, C. & Blaxton, T. (1998). Data mining solutions: methods and tools for solving real-world problems.

Wright, E.; Heckel, T.; Groenendijk, J.; Davies, J. & Boulton, M. (1997). Splicing features in maize streak virus virion-and complementary-sense gene expression. *The plant journal*, 12(6):1285--1297.

Yan, E. & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, 60(10):2107--2118.

Zaki, M. & Meira, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Zorzatto, C.; Machado, J. P. B.; Lopes, K. V.; Nascimento, K. J.; Pereira, W. A.; Brustolini, O. J.; Reis, P. A.; Calil, I. P.; Deguchi, M.; Sachetto-Martins, G. et al. (2015). Nik1-mediated translation suppression functions as a plant antiviral immunity mechanism. *Nature*, 520(7549):679--682.

Apêndice A

Arquivos suplementares do artigo 3

A.1 Arquivo suplementar 1

Tabela A.1. Valor do IG dos atributos utilizados para obter os modelos de classificação de gênero.

Atributo	Descrição do atributo	IG
full_length	Comprimento da sequÃªncia	0.968 (1°)
gc_R1	Proporção de guanina e citosina na região 1	0.626 (10°)
gc_R2	Proporção de guanina e citosina na região 2	0.626 (8°)
gc_R3	Proporção de guanina e citosina na região 3	0.626 (9°)
gc_R4	Proporção de guanina e citosina na região 4	0.575 (13°)
percent_a	Proporção de adenina no genoma	0.568 (14°)
percent_a_R1	Proporção de adenina na região 1	0.648 (5°)
percent_a_R2	Proporção de adenina na região 2	0.648 (6°)
percent_a_R3	Proporção de adenina na região 3	0.648 (7°)
percent_a_R4	Proporção de adenina na região 4	0.600 (11°)
percent_c	Proporção de citosina no genoma	0.480 (18°)
percent_c_R1	Proporção de citosina na região 1	0.666 (4°)
percent_c_R2	Proporção de citosina na região 2	0.666 (3°)
percent_c_R3	Proporção de citosina na região 3	0.666 (2°)
percent_c_R4	Proporção de citosina na região 4	0.590 (12°)
percent_g	Proporção de guanina in genome	0.443 (21°)
percent_g_R1	Proporção de guanina na região 1	0.548 (16°)
percent_g_R2	Proporção de guanina na região 2	0.548 (15°)
percent_g_R3	Proporção de guanina na região 3	0.548 (17°)
percent_g_R4	Proporção de guanina na região 4	0.460 (20°)
percent_t	Proporção de timina no genoma	0.467 (19°)
percent_t_R1	Proporção de timina no região 1	0.425 (23°)
percent_t_R2	Proporção de timina na região 2	0.425 (22°)
percent_t_R3	Proporção de timina na região 3	0.425 (24°)
percent_t_R4	Proporção de timina na região 4	0.393 (25°)

A.2 Arquivo suplementar 2

Tabela A.2. Valor do IG dos atributos utilizados para obter os modelos de classificação de genes

Atributo	Begomovirus		Mastrevirus		Becurtovirus		Curtovirus		Turncurtovirus		Eragrovirus		Alphasatellite/ Betasatellite	
Alanine	2.2292	9°	0.1793	27°	2.12592	18°	1.13407	29°	2.5275	13°	1.585	2°	0.752	21°
Arginine	2.0807	21°	0.4867	17°	2.08025	20°	1.18501	25°	2.4646	22°	1.355	29°	0.693	27°
Asparagine	2.0917	18°	0.2446	24°	2.22524	14°	1.62116	14°	2.5366	10°	1.585	15°	0.768	17°
Aspartic_acid	2.1647	15°	0.3426	19°	2.22813	12°	1.51427	17°	2.5850	2°	1.585	6°	0.732	23°
Chain	0.9448	32°	0.9993	2°	0.97095	35°	0.97951	35°	0.9183	35°	0.918	33°	0.818	7°
Cysteine	2.3211	4°	0.1087	31°	2.32193	11°	2.22356	4°	2.5850	4°	1.585	14°	0.879	4°
Glutamic_acid	2.0995	17°	0.3884	18°	2.32193	4°	1.38672	21°	2.4703	19°	1.585	5°	0.701	26°
Glutamine	2.1846	12°	0.8701	5°	2.03612	21°	1.43531	19°	2.5850	3°	1.585	16°	0.708	25°
Glycine	2.1818	13°	0.2207	25°	1.52992	32°	2.00611	8°	2.5366	9°	1.585	22°	0.761	18°
Histidine	2.2736	6°	1.3875	1°	2.08025	19°	2.06013	7°	2.5850	6°	1.585	26°	0.73	24°
Isoleucine	2.1487	16°	0.1793	28°	2.32193	3°	1.4756	18°	2.4925	17°	1.585	25°	0.883	3°
length	2.3724	2°	0.9218	3°	2.32193	1°	2.6775	1°	2.5850	1°	1.585	27°	0.911	1°
Leucine	2.2292	10°	0.9212	4°	2.12592	17°	1.13407	30°	2.5275	12°	1.585	28°	0.752	20°
Lysine	2.0816	20°	0.5060	15°	1.99371	24°	1.28882	23°	2.5366	11°	1.585	24°	0.733	22°
Methionine	2.2292	8°	0.1793	26°	2.12592	16°	1.13407	31°	2.5275	14°	1.585	20°	0.752	19°
Percent_a	0.9726	31°	0.8420	6°	1.69228	31°	1.06637	32°	2.4953	16°	1.585	11°	0.792	10°
Percent_a_R1	1.2305	23°	0.7445	10°	2.32193	8°	1.04738	33°	2.1865	32°	1.585	13°	0.789	11°
Percent_a_R2	1.2446	22°	0.7581	9°	2.32193	5°	1.00653	34°	2.0789	33°	1.585	8°	0.785	13°
Percent_c	1.1021	25°	0.2549	23°	1.92193	26°	1.95584	9°	1.9154	34°	0.918	34°	0.434	32°
Percent_c_R1	1.0560	28°	0.0889	32°	2.02261	22°	1.58388	16°	2.3711	29°	1.585	9°	0.504	31°
Percent_c_R2	0.9280	33°	0.0000	35°	2.00992	23°	1.58388	15°	2.3711	28°	1.585	3°	0.284	34°
Percent_g	0.6397	34°	0.7678	8°	1.92193	27°	1.13993	28°	2.2612	31°	1.355	30°	0.69	28°
Percent_g_R1	1.0724	27°	0.6789	11°	2.32193	7°	1.17317	27°	2.4702	20°	1.355	32°	0.794	9°
Percent_g_R2	0.9857	30°	0.5927	12°	2.22524	13°	1.17317	26°	2.3994	27°	1.355	31°	0.812	8°
Percent_genome	0.0000	35°	0.0000	34°	2.32193	10°	2.58167	2°	2.5850	7°	0.000	35°	0.91	2°
Percent_t	1.0758	26°	0.2847	22°	1.84381	29°	2.07685	6°	2.2825	30°	1.585	12°	0.684	29°
Percent_t_R1	1.0435	29°	0.3035	21°	1.92193	28°	1.36178	22°	2.4115	26°	1.585	10°	0.383	33°

Percent_t_R2	1.1225	24°	0.3214	20°	1.92193	25°	1.24468	24°	2.4148	25°	1.585	7°	0.255	35°
Phenylalanine	2.3292	3°	0.5687	14°	2.32193	6°	2.22172	5°	2.5044	15°	1.585	23°	0.771	16°
Proline	2.1921	11°	0.5919	13°	2.22524	15°	1.62669	13°	2.4549	24°	1.585	18°	0.843	5°
Serine	2.1661	14°	0.1328	30°	1.23836	34°	1.92694	10°	2.4550	23°	1.585	17°	0.559	30°
Threonine	2.0878	19°	0.1762	29°	1.81702	30°	1.86228	11°	2.4825	18°	1.585	19°	0.782	14°
Tryptophan	2.3025	5°	0.4969	16°	2.32193	9°	2.52714	3°	2.5850	8°	1.585	21°	0.776	15°
Tyrosine	2.3807	1°	0.0739	33°	2.32193	2°	1.73775	12°	2.5850	5°	1.585	4°	0.833	6°
Valine	2.2321	7°	0.8295	7°	1.37095	33°	1.38721	20°	2.4692	21°	1.585	1°	0.787	12°

Apêndice B

Trabalhos publicados ou em análise para publicação

A seguir estão relacionados os trabalhos publicados ou em análise para publicação, trabalhos estes resultantes desta dissertação:

- Capítulo de livro
  Título: Predição computacional de microRNAs em genomas (cap. 16).
  Livro: Introdução ao mundo dos microRNAs. 1ed.
  Editora: São Carlos.
  Autores: CERQUEIRA, F. R. ; MARQUES, Y. B. ; CARVALHO, T. F. M. ; SILVA, J. C. F. ; BASSO, M. F. ; MORAIS, G. L. ; CARVALHO, J. B.
  Ano: 2015.

- Apresentação de pôster
  Título: Rama: A machine learning approach for ribosomal protein prediction in plants.
  Autores: CARVALHO, T. F. M.; SILVA, J. C. F. ; FONTES, E. P. B. ; CERQUEIRA, F. R.
  Evento: X-Meeting 2015 - 11th International Conference of th AB3C + Brazilian Symposium of Bioinformatics, 2015.
  Local: São Paulo - Brasil.

Ano: 2015.

- Apresentação de pôster
  Título: Begomovirus Data Warehouse (BegomoDW): A definitive and integrated database for begomoviruses and related satellite virus
  Autores: José Cleydson F Silva; Marcos F Basso; Michihito Deguchi; Welison A Pereira; Roberto R Sobrinho3; Otávio JB Brustolini;, Pedro M P Vidigal; Alison T M Lima; Maximiller D L Costa; Francisco Murilo Zerbini; Fabio R Cerqueira1 and Elizabeth P B Fontes
  Evento: IPMB 2015 South America, 11th International Congrees of Plant Molecular Biology Iguazú Falls (Argentina - Brazil) October, 2015
  Local: Foz do Iguaçu - Brasil.
  Ano: 2015.

- Trabalho enviado para publicação (aguardado resposta)
  Título: Begomovirus Data Warehouse (BegomoDW): A definitive and integrated database for begomoviruses and related satellites.
  Autores: José Cleydson F Silva, Marcos F Bassoz, Thales F. M. Carvalho, Michihito Deguchiz, Welison A Pereiraz, Roberto R Sobrinhoz, Otávio JB Brustoliniz, Pedro M P Vidigalz, Alison T M Lima, Anésia A. Santosz, Maximiller D L Costaz,Francisco Murilo Zerbini, Fabio R Cerqueira e Elizabeth P B Fontes.
  Evento: CLEI 2016
  Ano: 2016.