THALES FRANCISCO MOTA CARVALHO

MÉTODOS DE MINERAÇÃO DE DADOS PARA EXTRAÇÃO DE CONHECIMENTO EM BIOINFORMÁTICA: APLICAÇÃO EM DADOS DE *GEMINIVIRUS* E PREDIÇÃO DE NOVAS PROTEÍNAS RIBOSSOMAIS

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
C3312m
2016

Carvalho, Thales Francisco Mota, 1990-
Métodos de mineração de dados para extração de conhecimento em bioinformática : aplicação em dados de *Geminivirus* e predição de novas proteínas ribossomais / Thales Francisco Mota Carvalho. – Viçosa, MG, 2016.
x, 72f. : il. (algumas color.) ; 29 cm.

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

1. Mineração de dados (Computação). 2. Bancos de dados. 3. Processamento de linguagem natural (Computação). 4. Bionformática. 5. *Geminivirus. 6.* Aprendizado de máquina. 7. Proteínas. I. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-graduação em Ciência da Computação. II. Título.

CDD 22. ed. 005.74

THALES FRANCISCO MOTA CARVALHO

MÉTODOS DE MINERAÇÃO DE DADOS PARA EXTRAÇÃO DE CONHECIMENTO EM BIOINFORMÁTICA: APLICAÇÃO EM DADOS DE GEMINIVIRUS E PREDIÇÃO DE NOVAS PROTEÍNAS RIBOSSOMAIS

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 este trabalho à minha amada mãe Dulce, ao meu irmão Tácio (in memoriam) e à minha namorada Vívian.

Agradecimentos

Agradeço a Deus por me amparar nos momentos difíceis, me dar força interior para superar as dificuldades e por sempre estar comigo.

À minha mãe Dulce, por sempre estar ao meu lado com o seu amor eterno e incondicional. Obrigado por ter me educado, pelos conselhos, incentivos, afeto, amizade e por ter se dedicado e contribuído para meu sucesso. Você muitas vezes renunciou aos seus sonhos para que eu pudesse realizar o meu, partilho com você a alegria deste momento. Se hoje cheguei aqui, tenho muito a agradecê-la. Obrigado ao meu irmão Tácio (*in memorian*) pelos momentos inesquecíveis, pela cumplicidade e amizade. "Tácio, a saudade é grande, mas o amor é para sempre. Mesmo não estando aqui, sinto sua presença e força me guiando. Sinta-se orgulhoso, pois, esta vitória é nossa!"

À minha família, a qual amo muito, pelo carinho, paciência e incentivo. Aos meus primos por todos os momentos que passamos juntos de pura amizade. À minha vó Perpétua, pelas orações e carinho. À família da minha namorada, por ter me recebido de coração aberto e sempre torcerem por mim.

À Universidade Federal de Viçosa pela oportunidade de realizar o curso. Ao Prof. Fábio, meu orientador e exemplo de profissional, por sua dedicação, paciência, confiança e por seus ensinamentos. À minha coorientadora Prof.ª Elizabeth, pela sua atenção, valiosas sugestões e por abrir as portas do seu laboratório, onde pude aprender muito. Agradeço também a todos os docentes do Departamento de Informática da UFV que de alguma forma contribuíram para o meu crescimento profissional. Ao funcionário Altino, pelas diversas assistências e agradáveis conversas.

Aos amigos que fizeram parte dessa caminhada, sempre me ajudando e incentivando. Obrigado Matheus, Renan Oliveira, Renan, Otávio e Fabiene. Obrigado todos os amigos de mestrado pelo companheirismo e apoio.

Ao Cleysinho, pelos seus ensinamentos, companheirismo e valiosa amizade. Sem você e sua dedicação, conhecimento e paciência, esse trabalho não seria possível. Essa nossa parceria pode ser comparada a dos *Hobbits Frodo Baggins* e *Samwise Gamgee*, que provaram que uma forte amizade e confiança um no outro, permitem superar enormes desafios e obter sucesso ao final de uma longa caminhada.

À Vívian, minha alma gêmea. Sua ajuda e apoio foram para mim de valor inestimável, e qualquer palavra que eu utilize seria insuficiente para agradecê-la com o devido merecimento. Obrigado por acreditar em mim quando eu achei difícil acreditar em mim mesmo. Obrigado por dizer, algumas vezes, o que eu realmente precisava ouvir, em vez do que eu queria que você dissesse. Tenho certeza que sem você eu não teria chegado até aqui. Entre as infinitas coisas que tenho que agradecê-la, um muito obrigado pela amizade, amor, companheirismo, lealdade, carinho e por cuidar de mim todos esses anos.

Por fim, o meu profundo e sentido agradecimento a todas as pessoas que contribuíram para a concretização desta dissertação, estimulando-me intelectualmente e emocionalmente.

"O saber a gente aprende com os mestres e os livros. A sabedoria, se aprende é com a vida e com os humildes."(Cora Coralina)

Sumário

Resumo

Carvalho, Thales Francisco Mota, M.Sc., Universidade Federal de Viçosa, julho de 2016. **Métodos de mineração de dados para extração de conhecimento em bioinformática: aplicação em dados de *Geminivirus* e predição de novas proteínas ribossomais**. Orientador: Fábio Ribeiro Cerqueira. Coorientadora: Elizabeth Pacheco Batista Fontes.

A mineração de dados (DM, do inglês *data mining*) é um processo de descoberta de padrões que permite extrair informação e conhecimento em grandes volumes de dados. Suas principais técnicas se baseiam em predição, classificação e agrupamento (*clustering*). Estas técnicas têm sido utilizadas na bioinformática para classificar o perfil de expressão gênica, encontrar padrões em sequências de DNA, avaliar a estrutura do dobramento de proteínas, entre outras aplicações. Neste trabalho, avançadas técnicas de DM foram aplicadas para o desenvolvimento de um *Data Warehouse* específica para geminivírus (*geminivirus.org*), a fim de auxiliar na organização, correção e normalização de dados referentes a geminivírus. Neste *Data Warehouse* também foram propostas metodologias baseadas em regras e aprendizado de máquina (ML) que classificam as sequências de DNA e seus genes. A família *Geminiviridae* é composta por pequenos vírus de DNA circular de fita simples que infectam uma grande variedade de plantas e causam sérios danos econômicos ao redor do mundo. O aprimoramento da amplificação do DNA viral e de técnicas de sequenciamento permitiram um enorme crescimento de dados em banco de dados públicos. Simultaneamente, ocorreu o crescimento no volume de publicações relacionadas a esta família. Desta forma, numa segunda linha de trabalho surgiu a necessidade de aplicar as técnicas de DM, seguindo o processo de

KDD (*knowledge-discovery in databases*) para extrair informações desses dados. Além disso, técnicas de Processamento de Linguagem Natural (NLP) foram utilizadas para extrair informação em resumos de artigos relacionados a geminivírus. Assim, o acervo científico pode ser explorado de maneira contextualizada. Finalmente, uma terceira frente de trabalho em mineração de dados foi empreendida, desta vez direcionada à descoberta de proteínas ribossomais. Pesquisas recentes têm demonstrado que plantas suprimem o mecanismo global de tradução como uma estratégia de imunidade antiviral. Entretanto, poucas proteínas ribossomais são mencionadas a integrarem vias do mecanismo de defesa das plantas. As proteínas ribossomais (RPs) desempenham um papel fundamental em células vivas, pois são o principal componente dos ribossomos. Além disso, estas proteínas estão envolvidas em vários processos fisiológicos e patológicos. Assim, foi desenvolvido um método de aprendizado de máquina capaz de identificar novas proteínas ribossomais, designado Rama. O Rama utiliza abordagens inovadoras em relação aos métodos computacionais atualmente existentes. Em experimentos *in silico*, o Rama obteve resultados médios de precisão, acurácia, sensitividade e especificidade de 0.9203, 0.9214, 0.9214 e 0.8236, respectivamente. Ademais, duas proteínas não caracterizadas foram preditas como RPs pelo Rama e experimentos *in vitro* confirmaram a veracidade do resultado, ao passo que as metodologias atuais não conseguem lograr o mesmo sucesso.

Abstract

Carvalho, Thales Francisco Mota, M.Sc., Universidade Federal de Viçosa, July of 2016. **Data mining methods for knowledge extraction in bioinformatics: Application on Geminivirus data and prediction of new ribosomal proteins** Adviser: Fábio Ribeiro Cerqueira. Co-advisor: Elizabeth Pacheco Batista Fontes.

Data mining (DM) is a pattern discovery process that can extract information and knowledge in large volumes of data. Its main techniques are based on prediction, classification, and clustering. These techniques have been used in bioinformatics to identify gene expression profiles, find patterns in DNA sequences, evaluate protein folding structure, among other applications. In this work, advanced techniques of DM were applied to the development of a specific Data Warehouse for geminiviruses (*geminivirus.org*) to assist in organization, correction, and normalization of data related to geminivirus. In this Data Warehouse, we also propose methodologies based on rules and machine learning (ML) to classify DNA sequences and their genes. The *Geminiviridae* family consists of small circular single-stranded DNA viruses which infect a wide variety of plants and cause serious economic losses wordwide. Improvements in amplification of viral DNA and sequencing techniques have led to an enormous growth of public databases. Thus, in a second endeavor in this work, we realized the need to apply DM techniques, following the process of KDD (knowledge-discovery in databases), to extract yet-unknown information. Furthermore, natural language processing techniques (NLP) were used to extract information in abstracts of paper related to geminivirus. In this way, the scientific literature can be explored in a contextualized manner. Finally, a third effort using

data mining approaches was carried out, this time directed to the identification of new ribosomal proteins. Recent research has shown that plants suppress the overall mechanism of translation as a strategy for antiviral immunity. However, few ribosomal proteins are referred to integrate pathways of plant defense mechanisms. Ribosomal proteins (RPs) have a fundamental role in living cells, as they are the main component of ribosomes. Furthermore, these proteins are involved in various physiological and pathological processes. Therefore, we developed a ML method to identify new ribosomal proteins, called Rama. Rama uses innovative approaches in comparison to currently existing computational methods. In *in silico* experiments, Rama presented average results of precision, accuracy, sensitivity, and specificity of 0.9203, 0.9214, 0.9214, and 0.8236, respectively. In addition, two proteins not yet characterized were predicted as RPs by Rama, whereas other methods could not achieve the same success. *In vitro* experiments confirmed the veracity of our result.

Capítulo 1

Introdução

1.1 Introdução geral

A mineração de dados (DM, do inglês *data mining*) é o processo de descoberta de padrões fundamentais, interessantes e inovadores (Zaki & Meira, 2014). Entre as diversas funcionalidades apresentadas na literatura para as técnicas de DM, pode-se destacar: predição, classificação, segmentação, agrupamento (*clustering*), visualização e otimização (Westphal & Blaxton, 1998)(Weis & Indurkhya, 1999)(Mena, 1999).

A vasta possibilidade de aplicação da DM permite que diferentes áreas de conhecimento a utilizem para resolução de problemas específicos de cada área. Por exemplo, na área da saúde pode-se utilizar DM para predição de risco de doença cardíaca (Purusothaman & Krishnakumari, 2015), análise de diagnóstico e tratamento de câncer de Mama (Lu et al., 2015). A DM pode auxiliar também na identificação de necessidades da saúde dos usuários, bem como na organização dos serviços de saúde para suprir tais necessidades (Souza & Zaia, 2015). Já na área comercial, pode ser utilizada para detectar padrões de comportamento do cliente para formulação de estratégias de marketing, vendas e suporte ao cliente (Berry & Linoff, 1997).

Na bioinformática, as técnicas de DM são amplamente utilizadas auxiliando os pesquisadores a analisar e entender um grande volume de dados, tais como dados gerados por sequenciadores de nova geração. A DM também é empregada de

diversas formas em dados biológicos como, por exemplo, para observar proteínas, selecionar expressão gênica baseado em microarray, classificar o perfil de expressão gênica, encontrar padrões em sequências de DNA, avaliar a estrutura em dobramento de proteínas e descobrir mecanismos genéticos que ainda não são entendidos (Frank et al., 2004)(Wang et al., 2005)(Ma et al., 2014).

Um dos fatores que permite a DM ser tão flexível e com diversas funcionalidades é a sua vasta quantidade de técnicas, as quais permitem diferentes processos de descobertas de conhecimento. Para auxiliar no entendimento dessa grande quantidade de métodos, Maimon & Rokach (2005) apresentam uma taxonomia que ilustra a inter-relação e agrupamento de tais métodos (Figura 1). Neste agrupamento, é interessante distinguir a DM em dois principais tipos: orientado a verificação (verifica a hipótese do usuário) e orientado a descoberta (encontra novas regras e padrões de forma autônoma) (Maimon & Rokach, 2005).

Uma forma de descobrir padrões utilizando DM é aplicar métodos de aprendizado de máquina (ML, do inglês *machine learning*), os quais auxiliam na classificação de padrões. Métodos não supervisionados, como agrupamento (clusterização) e sumarização, também podem ser usados para encontrar padrões. A DM também integra o processo de extração de conhecimento (KDD, do inglês *knowledge-discovery in databases*)(Maimon & Rokach, 2005), em que auxilia na padronização, na descoberta e apresentação da informação ao usuário final.

Tendo em vista a diversidade e capacidade das técnicas de DM para processar informações biológicas e a fim de auxiliar pesquisadores a solucionarem algumas questões que foram identificadas como relevantes, neste trabalho, foi proposta a aplicação de DM em dados derivados de geminivírus e proteínas ribossomais. Tais técnicas foram utilizadas para: 1) recuperação, processamento e normalização de sequências de geminivírus; 2) análise de assuntos mais abordados em artigos científicos sobre geminivírus; e 3) classificação de proteínas ribossomais baseada em ML.

A família *Geminiviridae* é uma família de vírus com DNA circular de cadeia simples (ssDNA) que infecta uma grande variedade de plantas dicotiledôneas e monocotiledôneas. Essa família é composta por sete gêneros: *Becurtovirus*, *Begomovirus*, *Curtovirus*, *Eragrovirus*, *Mastrevirus*, *Topocuvirus* e *Turncurtovirus*. Tais gêneros estão relacionados a sérios problemas, por exemplo, o *begomovirus* é am-

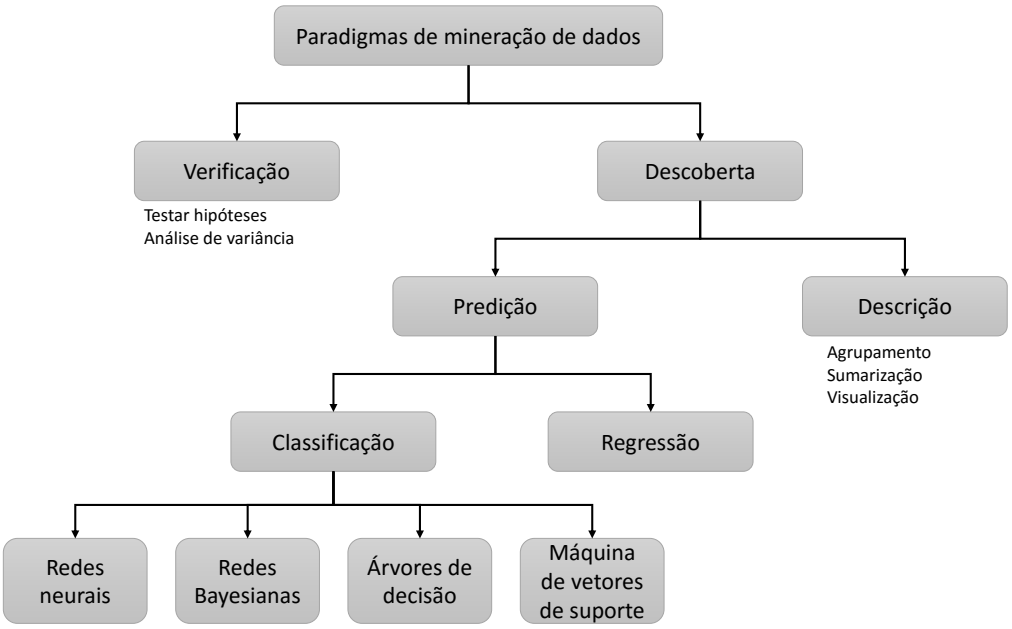

Figura 1.1. Taxonomia do processo de *Data Mining* (adaptado de Maimon & Rokach (2005))

plamente distribuído em todo o mundo e causa doenças de alto impacto econômico sobre uma ampla gama de culturas importantes (Prasanna & Rai, 2007).

Devido à importância e grande quantidade de pesquisa envolvendo a família *Geminiviridae*, uma ampla quantidade de DNA e de proteínas têm sido anotada e armazenada em bancos de dados do NCBI (*National Center for Biotechnology Information*). Esta grande quantidade de genomas anotados e a falta de regras mais rígidas para armazenar tais sequências resultaram em um banco de dados de geminivírus do NCBI com pouca padronização, dados incompletos e até mesmo inconsistentes (e.g., nome das proteínas que compõem os vírus).

Para solucionar tais problemas, viu-se a possibilidade de usar as técnicas de DM juntamente com o KDD para criar um *Data Warehouse* específico para geminivírus, podendo, então, padronizar e auxiliar a descoberta de novos conhecimentos. Sendo assim, criou-se o *geminivirus.org*, o qual é um *Data Warehouse* capaz de armazenar, corrigir e padronizar as informações do NCBI utilizando con-

juntos de regras e ferramentas de bioinformática. Metodologias baseadas em regras padronizam as sequências na origem de replicação, ajustam o posicionamento dos genes e relacionam suas informações (hospedeiro de origem, localização geográfica e data de coleta). Além disso, por meio do *geminivirus.org* é possível realizar predições/classificações utilizando genomas completos a partir de técnicas de ML, como pode ser visto no Capítulo 2 descrito mais à frente.

Para cada sequência de geminivírus anotada no NCBI, o *geminivirus.org* realiza uma série de verificações para conferir se o genoma possui as características de cada gênero de geminivírus descritas na literatura, por exemplo, tamanho mínimo e máximo da sequência. Posteriormente, é realizada uma padronização de informação, a qual padroniza a origem de replicação do genoma e o nome dos genes. Tais informações corrigidas são armazenadas em uma base de dados relacional estruturada para armazenar de forma eficiente estes dados. Os vírus armazenados são submetidos às técnicas de enriquecimento de informação, como descoberta de coordenadas geográficas e classificação dos genes utilizando métodos de ML. Também é realizada uma busca manual por informações contidas nos artigos que descrevem a anotação das sequências do vírus. O *geminivirus.org* também recupera e armazena os resumos das publicações referentes a geminivírus armazenadas no PubMed[1] para relacionar os vírus com suas respectivas publicações.

Após a criação do acervo de resumos de publicações referentes a geminivírus, percebeu-se um grande potencial em explorar técnicas de DM para encontrar informações relevantes derivadas do relacionamento entre todos os resumos. Com isso, criou-se uma ferramenta web baseada em métodos de processamento de linguagem natural (NLP, do inglês *natural language processing*) para analisar as palavras contidas nesses resumos, como pode ser visto no Capítulo 3.

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 visualizar 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 na rede

[1]PubMed é um motor de busca de livre acesso à base de dados MEDLINE de citações e resumos de artigos de investigação em biomedicina.

de palavras. O *betweenness centrality* baseia-se no número de caminhos mais curtos que passam através de um vértice, onde um elevado *betweenness* pode indicar vértices que definem grupos. O *Degree centrality* é obtido a partir do número de arestas que liga um vértice com outros vértices. Já o *PageRank* é bastante utilizado para classificação de páginas web com base em um algoritmo de propagação de peso (Yan & Ding, 2009). Tais métricas podem indicar quais são as palavras com maior peso na rede e como elas interagem entre si dentro do contexto acadêmico em geminivírus. O grafo permite verificar como um determinado estudo está sendo abordado em geminivírus, 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 existem fortes estudos focados na recombinação e mutação de tais vírus no contexto de evolução em geminivírus.

Para facilitar a utilização e interpretação do grafo e pesos gerados, foi desenvolvido uma ferramenta web em que o pesquisador informa um conjunto de palavras e recebe como resposta um grafo interativo juntamente com uma tabela informando os pesos de cada palavra nos estudos de geminivírus. Essa ferramenta também permite comparar os grafos e pesos das palavras de conjuntos de resumos de artigos dos patógenos dengue, Ebola, H1N1 e Zika vírus. Tal comparação pode ser analisada interagindo com dois grafos um ao lado do outro, com os pesos tabulados e um *heat map* (mapa de calor) mostrando a frequência das palavras nos resumos dos patógenos.

Observando os resultados alcançados na aplicação de DM na criação do *geminivirus.org* e diversos trabalhos que utilizam o DM para solucionar problemas de bioinformática, surgiu o interesse em aplicar métodos de ML para encontrar novas proteínas ribossomais em plantas. Tais proteínas ribossomais (RPs) desempenham um papel fundamental dentro de todos os tipos de células, uma vez que são o principal componente dos ribossomos. O ribossomo é uma importante maquinaria celular responsável pela síntese de diferentes proteínas (Alberts et al., 1995). Pesquisas recentes tem demonstrado que plantas suprimem o mecanismo global de tradução como uma estratégia de imunidade antiviral (Zorzatto et al., 2015). Além disso, sabe-se que RPs estão envolvidos em vários processos fisiológicos e pa-

tológicos, e podem agir como um efetor da resposta antiviral em plantas (Rocha et al., 2008).

Atualmente, a principal ferramenta utilizada para identificar proteínas ribossomais é o InterProScan. O InterProScan apresenta anotações funcionais e realiza análises que permitem identificar, sempre que possível, a ontologia de genes associados a termos do *Gene Ontology Consortium* (Mitchell et al., 2015). Por mais robusta que esta ferramenta seja, ela está limita a identificar proteínas ribossomais que sejam parecidas com as proteínas já conhecidas, ou seja, esta ferramenta nem sempre é capaz de identificar novas proteínas ribossomais, uma vez que baseiam-se em homologia de sequência.

Tendo em vista a importante funcionalidade das RPs e limitações da principal ferramenta de identificação, surgiu o interesse em verificar os estudos de tais proteínas utilizando o grafo de texto descrito acima. O grafo de texto mostrou que elas não são amplamente estudadas, já que a palavra *ribosome* e *ribosomal* possui poucas conexões e baixo peso nas métricas utilizadas. Sendo assim, houve uma motivação ainda maior para desenvolver novas técnicas de predição de RPs.

Para encontrar novas RPs, desenvolveu-se uma metodologia que utiliza os métodos de ML *Multilayer Perceptron* (MLP), *Random Forest* (RF) e *Support Vector Machine* (SVM) para criar modelos de classificação capazes de classificar sequências proteicas em RPs ou proteínas não ribossomais (NRPs). Estes modelos foram treinados apenas com características fundamentais dos aminoácidos. Tal metodologia deu origem à ferramenta denominada Rama, como pode ser visto no Capítulo 4 da presente dissertação.

O Rama é composto por duas etapas: (i) classificação de RPs e NRPs e (ii) classificação de RPs e proteínas histonas (HPs). Este último passo é necessário, pois, RPs e HPs possuem uma certa semelhança. Na primeira etapa, são selecionados 6 modelos de classificação (de um total de 36 modelos) treinados com RPs/NRPs de cinco espécies de plantas (*Zea mays*, *Oryza sativa*, *Arabidopsis thaliana*, *Solanum lycopersicum* e *Glycine max*) e um fitoplâncton (*Ostreococcus lucimarinus*). Estes seis modelos podem variar dependendo da espécie de origem da proteína informado pelo pesquisador. Cada modelo selecionado realiza a classificação da proteína desconhecida em RP ou NRP e a classificação final será dada por um processo de votação em que a classe majoritária dada pelos modelos será a

vencedora. Se a proteína for classificada como RP ela será submetida ao segundo passo do Rama. Da mesma forma do primeiro passo, no segundo passo a proteína é submetida em outros seis modelos, porém estes modelos são ajustados para diferenciar RPs de HPs. Caso o consenso desses seis outros modelos classifiquem a proteína como RP ela será de fato considerada um proteína ribossomal, caso contrário, ela será considerada NRP.

Em testes *in silico*, o Rama obteve resultados médios de precisão, acurácia, sensitividade e especificidade de 0.9203, 0.9214, 0.9214, e 0.8236, respectivamente. O Rama foi capaz de indicar duas novas proteínas ribossomais, as quais o InterProScan não consegue realizar nenhuma predição. Ambas foram validadas em experimentos *in vitro*.

Neste trabalho, foram desenvolvidas três ferramentas (online), baseadas em técnicas de DM, que visam contribuir para resolução de problemas que foram identificados como pouco explorados mas que apresentam clara motivação para novos estudos científicos. Para cada ferramenta desenvolvida, foi redigido um artigo científico, os quais são apresentados nos próximos capítulos. O segundo capítulo é composto pelo artigo que descreve detalhadamente como são realizados a padronização e o ganho de informação no *geminivirus.org*. O terceiro capítulo apresenta a metodologia seguida para construção do grafo de texto e como utilizá-lo para verificar um contexto de estudo em geminivírus. Por fim, no quarto capítulo, é descrito o método Rama, onde são apresentados sua metodologia e seus resultados *in silico* e *in vitro*.

1.2 Estrutura do trabalho

Esse primeiro capítulo apresentou uma introdução geral dos trabalhos realizados nesta dissertação. Nos próximos capítulos são apresentados três artigos científicos. Cada artigo apresenta diferentes métodos de *data mining* para extração de conhecimento em bioinformática. São eles:

(a) Capítulo 2: Geminivirus Data Warehouse: A database enriched with machine learning approaches. Artigo a ser submetido para a revista Bioinformatics Oxford;

(b) Capítulo 3: Extração de informação entre vírus de planta e importantes patógenos virais de humanos usando abordagens de processamento de linguagem natural. Artigo a ser submetido para a revista Bioinformatics Oxford;

(c) Capítulo 4: Rama: A machine learning approach for ribosomal protein prediction in plants. Artigo a ser submetido para a revista Bioinformatics Oxford.

No quinto e último capítulo é apresentado uma conclusão geral.

Capítulo 2

Geminivirus Data Warehouse: A database enriched with machine learning approaches

Artigo científico a ser submetido para a 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 3

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

Artigo científico a ser submetido para a 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$_r eference.jsp$.

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

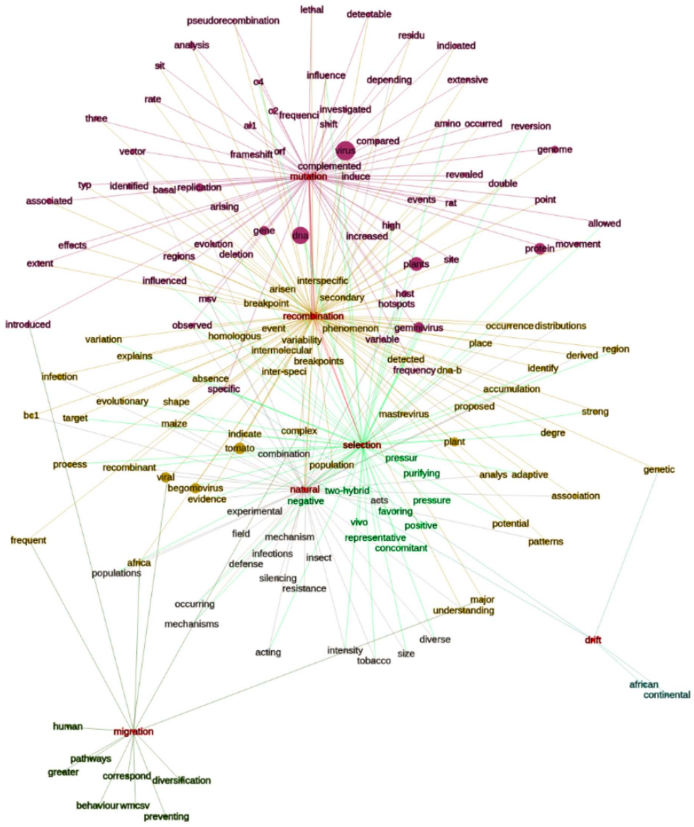

Figura 1 Grafo de palavras chave que representam os mecanismos evolutivos Apresenta um grafo de palavras é formado por meio de palavras chave. As palavras chave estão mascadas em vermelho. Cada palavra chave está conectada a nós vizinhos por meio de arestas com cores diferentes.

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.

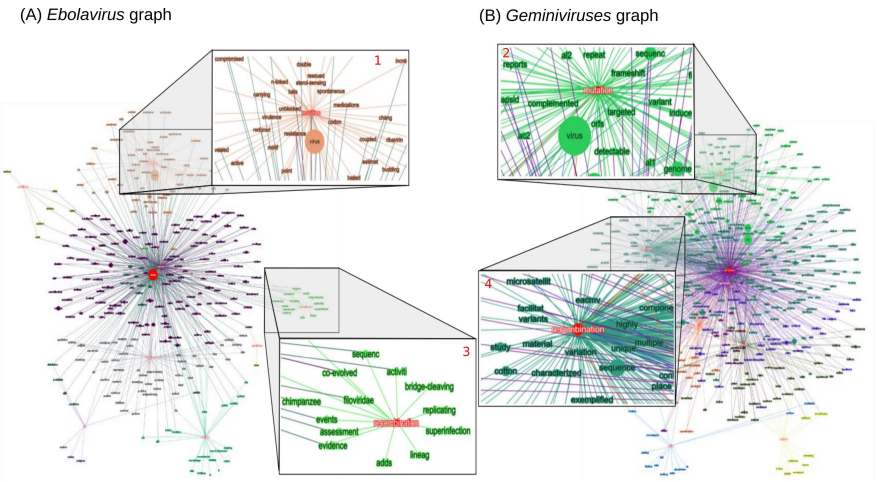

Figura 2 : Ilustração da comparação entre grafos de geminivírus e vírus ebola A figura apresenta dois grafos: o primeiro grafo (A) gerado pelo conjunto de dados de vírus ebola e o segundo grafo (B) gerado pelo conjunto de dados de geminivírus. Os quadros 1 e 2 apresentam as conexões do nó/token que representam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

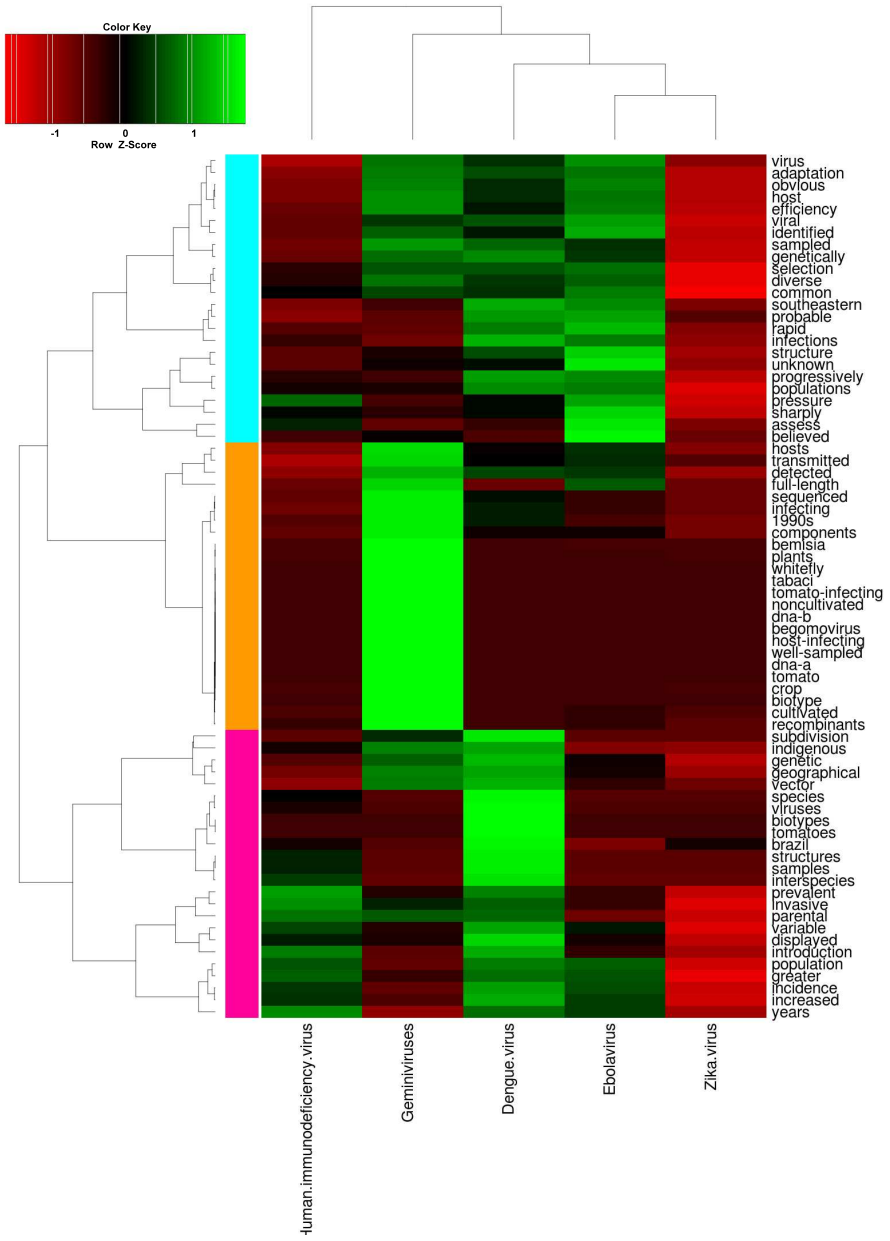

Figura 3 Mapa de calor para a clusterização de todas as palavras de um resumo O agrupamento apresenta grupos mais próximos de palavras coloridos em azul, laranja e rosa. As cores verdes apontam as palavras com maior grau no grafo, enquanto que as cores vermelhas mostram aquelas que menos ocorrem. As 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 4

Rama: A machine learning approach for ribosomal protein prediction in plants

Artigo científico a ser submetido para a revista Bioinformatics Oxford.

METHODOLOGY

Rama: A machine learning approach for ribosomal protein prediction in plants

Thales FM Carvalho[1†], José CF Silva[1,2†], Iara P Calil[2], Elizabeth PB Fontes[2] and Fabio R Cerqueira[1*]

*Correspondence:
fabio.cerqueira@ufv.br
[1]Departamento de Informática, Universidade Federal de Viçosa, 36570-900, Viçosa, Brazil
Full list of author information is available at the end of the article
[†]Carvalho and Silva contributed equally to this work

Abstract

Background: Ribosomal proteins (RPs) play a fundamental role within all type of cells, as they are major components of ribosomes without which no translation of mRNAs would be possible. Furthermore, these proteins are involved in various physiological and pathological processes. These facts motivate advanced studies for the identification of unrevealed RPs. In this work, we propose a new *in silico* method for the prediction of RPs, termed Rama, based on machine learning (ML) techniques, with a particular interest in plants. To perform an effective classification, Rama uses a set of fundamental features of the amino acid side chains and applies a two-step procedure to classify a set of proteins with unknown function as RPs. First, the models created from the datasets of RPs/non-RPs are applied; second, the proteins classified as RPs are given as input to the models created from the datasets of RPs/histones. Only the proteins considered as positives in these two classification filters are reported as RPs.

Results: For the experiments, we used data from six different plant species, retrieved from Phytozome database v10. Our results showed that Rama could achieve precision, accuracy, sensitivity, and specificity of 0.92, 0.92, 0.92, and 0.82, on average, respectively. Furthermore, our models classified several unannotated proteins as RPs with very high probability. Some selected proteins of *Arabidopsis thaliana* were validated in biological experiments.

Conclusions: In contrast to current computational methods for functional genomics, such as InterProScan, which are limited to the identification of conserved proteins, the ML-based Rama method was capable of predicting new RPs with high probability using only amino acid features. Furthermore, the classification of proteins with ML methods was much faster compared with methods that access multiple databases.

Keywords: machine learning; ribosomal protein; protein function prediction

Background

Ribosomal proteins (RPs) of different sizes associated with ribosomal RNAs (rRNAs) makes up the ribosome, an important cellular machinery responsible for synthesizing proteins [1]. The ribosome structure comprises two subunits: The small subunit (SSU) and the large subunit (LSU), the latter being approximately twice as large as the former. The association of ribosomes with messenger RNAs (mRNAs) allows the translation of these molecules, resulting in synthesis of proteins, which are vital for the cellular activities. Moreover, RPs are involved in several physiological and pathological processes. For example, RPs have been shown to trigger a suppression pathway for p53 tumors as a response to ribosomal stress [2]. Other important roles of RPs inside the cell have been also reported, such as effectors of

antiviral response in plants [2]. Identifying new ribosomal proteins may contribute to understanding how the ribosome works and discover new biological functions. This study is particularly focused on plant ribosomal proteins.

In several projects in genomics, metagenomics, and proteomics, the functional annotation of genes and proteins is aided by a large number of computational tools in order to identify specific functions or domains, such as active sites, functional domains, gene families, physical structures, or subcellular localization. The software InterProScan v.5 is one of the major computational tools in functional genomics and is interconnected with the main databases to carry out functional analysis of proteins [3]. InterProScan searches for protein signatures using methods included in two main modalities. The first one includes algorithms, such as TMHMM, SignalP, and Phobius, which work individually to check for any particular characteristic in the protein, e.g., membrane segments and signal peptide regions. The methods in the second modality use algorithms such as BLAST and hidden Markov models (HMM) to carry out searches for sequence alignments and perform protein associations, using a broad range of databases, providing the user with the result of a post-processing that matches output information produced by those methods. InterProScan features global functional annotations and its analyses enable identifying, whenever possible, gene ontology by associating terms from the Gene Ontology Consortium [4].

Although InterProScan is a robust tool to classify proteins in terms of function, its technique is limited to the databases used. This impairs the prediction of proteins that have no phylogenetic conservation or unknown functional domains. As a result, creating new computational methods to predict novel proteins is needed as a complement to the comparative analysis. We are particularly interested in ribosomal proteins in plants. In the present research, we propose a new *in silico* ribosomal protein prediction technique called Rama based on machine learning (ML) methods. Our approach uses characteristics common to any protein of an organism to build classification models. In this study, six organisms were selected to make up the training sets (TSs): Two species of monocotyledons: *Zea mays* (*Z. mays*) and *Oryza sativa* (*O. sativa*); three species of dicotyledons: *Arabidopsis thaliana* (*A. thaliana*), *Solanum lycopersicum* (*S. lycopersicum*), and *Glycine max* (*G. max*); and one species of phytoplankton: *Ostreococcus lucimarinus* (*O. lucimarinus*). As both RPs and histone proteins (HPs) display binding affinity to nucleic acids, i.e., they present similar amino acid properties, two training sets were created for each species, one containing RPs and non-ribosomal proteins (NRPs), and the other containing RPs and HPs. The amino acid sequences of these species were obtained from the repository Phytozome v.10 [5]. Once the training sets were determined, three ML algorithms were applied to build the predictive models: Multilayer perceptron (MLP), random forest (RF), and support vector machines (SVM). In the case of SVM, we used the sequential minimal optimization (SMO) algorithm, commonly applied for the learning process in the SVM approach [6][7]. Thousands of tests (cross-validation and inter-species) were run in order to estimate the best input parameters for each algorithm before the generation of the final ML models.

The prediction method hereby proposed performs two classification steps. Initially, sequences of proteins whose role is unknown are given as input to six models

trained with ribosomal and non-ribosomal proteins (RPs/NRPs), one model for each species. In the second step, the positively classified sequences undergo a new classification step using six other models trained with ribosomal and histone proteins (RPs/HPs), also one model for each species. The sequences classified positively in both steps are considered ribosomal proteins. Our approach is capable of revealing new components of the translation machinery, and can be used as an important complementary procedure in conjunction with methods based on homology using protein family databases.

The results showed that Rama reaches mean values for precision, accuracy, sensitivity, and specificity of 0.92, 0.92, 0.92, and 0.82, respectively. When Rama was applied on a set of unknown proteins of *A. thaliana*, the models identified some proteins with very high probability as RPs. Among those, the two top-ranked sequences were selected for *in vitro* validation, using a nucleic acid binding assay. This experiment demonstrated that the two predicted proteins exhibited biochemical activity of ribosomal proteins as they strongly bound RNA but not double-stranded DNA (dsDNA). These results also confirmed the efficiency of our filter to distinguish RPs from histones, which bind strongly to dsDNA. Interestingly, using the same sequences as input, InterProScan could not find any evidence for their association with the ribosome. Collectively, these results indicate that the ML-based Rama method is effective to predict new ribosomal proteins, and might be an important complement to the classical homology-based methods. Another important observation is that Rama's run time is approximately 600 times quicker compared to InterProScan.

Rama is available under the URL: inctipp.bioagro.ufv.br:8080/Rama/.

Materials and Methods

Training Sets

In order to create the training datasets, several amino acid sequences of proteins with their functional annotations were acquired in 2015 from the Phytozome v.10 database of five plant species and one phytoplankton species. Among the plant species, the sequences were retrieved from two monocotyledon species (*O. sativa* and *Z. mays*) and three dicotyledon species (*G. max*, *A. thaliana*, and *S. lycopersicum*). The phytoplankton species chosen was *O. lucimarinus*. The plant species were chosen for being widely studied. The phytoplankton (*O. lucimarinus*) was chosen as a common ancestral of the other species selected. Overall, thousands of proteins (amino acid sequences) were obtained, namely, 35,386 of *A. thaliana*, 88,647 of *G. max*, 49,061 of *O. sativa*, 7,796 of *O. lucimarinus*, 34,727 of *S. lycopersicum*, and 88,760 of *Z. mays*.

After data acquisition from Phytozome, filters were applied to remove unknown proteins with no functional annotation, and proteins that have not been characterized.

After the filter step, two training sets were generated for each species, one containing RPs (positive class)/NRPs (negative class) and the other containing RPs (positive class)/HPs (negative class). As expected, the resulting training sets of RPs/NRPs were extremely imbalanced, i.e., the number of NRPs was much greater than the number of RPs. Imbalanced TSs can result in models biased to the majority class, i.e., the minority classes, which are often the ones of most interest, are

classified with poor accuracy. As a consequence, we adjusted the sampling size from the NRPs datasets so that the final number of NRPs was three times greater than the number of RPs. According to our tests (not shown), this proportion was enough to avoid the bias problem, while not losing important information in the negative class. The number of proteins used in the RPs/NRPs and RPs/HPs training sets is presented in Tables 1 and 2, respectively.

Using this approach, 12 training sets, which included one dataset of RPs/NRPs and one dataset of RPs/HPs for each of the six species, were generated to be used as TSs for building the models of our two-stage classification procedure.

Feature Extraction

Amino acids can be categorized by chemical properties of their side-chains, such as polarity, net charge, etc. [8]. Such characteristics are the same in all organisms. As a result, the protein attributes used in this study originated from such chemical properties of amino acids. Specifically, the following proportions are collected for each protein: Amino acids that are aromatic (aromatic), negatively charged (negatively_charged), non-polar aliphatic (nonpolar_aliphatic), polar uncharged (polar_uncharged), positively charged (positively_charged), and hydrophobic (hydrophobic). The molecular mass (molecular_mass), volume (volume), and amount of amino acids (amount_of_amino_acids) were also measured. Each attribute contributes to distinguishing RPs from other proteins. For instance, RPs exhibit greater values for the attribute positively_charged compared to RNPs, because a positive charge facilitates ribosome binding to RNA (which has negative molecular charge).

To measure the importance of each of those attributes in RPs prediction, we used Information Gain (IG) [9]. This method assesses an attribute by measuring the gain in information in relation to the class and is defined by $IG(Attribute) = Entropy(Class) - Entropy(Class|Attribute)$. Entropy is given by $-\sum p_i log_2 p_i$, where p_i is the probability of class. Therefore, the best attributes will be the ones leading to the highest declines of entropy regarding the class.

Machine Learning Algorithms

In order to create models capable of predicting RPs for the species chosen in this study, some of the most well known ML algorithms were tested: Multilayer Perceptron (MLP), Random Forest (RF), and Sequential Minimal Optimization (SMO). These algorithms are implemented in the software Weka v3.7.11[10], used in this work.

MLP is likely the most commonly used architecture of the so-called artificial neural network (ANN) approach [6]. MLP features three types of layers containing artificial neurons. The input layer has one neuron for each attribute of the training instances. This layer receives the values of each attribute to be computed by the network. The final computed value will be available in the output layer. In a binary classification, this layer may have one or two neurons. Between the input and output layers, there is one or more intermediate layers containing one or more neurons. The links between the neurons do not form cycles, i.e., each neuron in the input layer and in the intermediate layers can be connected only to the neurons in the layer

immediately ahead (feedforward connections). Those connections have associated weights that are determined in the learning process and define the final value yielded by the network for a vector of attribute values provided as input. To establish the connection weights, the most popular learning method used for MLPs is called backpropagation. In this case, the difference found between the network output and the value observed in the training set for a given training instance is used to change the weights of connections in the opposite direction as the network's computation, i.e., from the output layer to the input layer.

The RF method is a type of ensemble approach, i.e., the final classification of an instance is performed from the majority vote of several models [11]. In the case of RF, the models are decision trees, each one built from a sample of the original training set. Each sample is generated taking a subset of the attributes. This method has been widely used in recent bioinformatics research with satisfactory results [12][13].

SMO is a widely applied training technique for the SVM approach. The SVM method aims to find a maximum margin hyperplane so as to minimize overfitting. The hyperplane is defined from the solution of a quadratic optimization problem. When the data are not linearly separable, a kernel function is used to perform an implicit transformation of the space of attributes into a higher dimensional space in which the instances of different classes can be separated by a hyperplane [6].

Those algorithms are used to create classification models from the training sets of RPs/NRPs and RPs/HPs. Therefore, when trained with RPs/NRPs, the learning models classify a protein as either ribosomal or non-ribosomal (step 1 of Rama), and when trained with RPs/HPs, they classify a protein as either ribosomal or histone-like (step 2 of Rama).

In order to choose the best ML algorithms and adjust the respective parameters, thousands of experiments were performed using different combinations of methods and parameter values. For each parameter setting for each algorithm, two types of experiments were carried out: One using a single species (both for training and for testing) and the other mixing the six species, i.e., the model is trained with one species and then tested with a different species (inter-species). In the experiments with a single species, 10-fold cross validation was used to measure the performance of the algorithms [14]. All these training/test processes were performed separately for RPs/NRPs datasets and for RPs/HPs datasets.

At each training/test experiments, the MLP, RF, and SMO algorithms were used in 6,000 runs, 40,931 runs, and 2,560 runs, respectively. Each run was performed with different parameter values. Additional file 1 shows the variations of parameters for the construction of all these models. After the execution of the algorithms, those that obtained the best precision and accuracy (Equation 1 and 2) in the inter-species experiments and also in the cross validation tests were chosen.

Thus, at the end of all experiments, 72 algorithms with suitable parameter values were chosen (see Additional file 2). Each algorithm gave rise to a distinct model for each training/test possibility from the RPs/NRPs and RPs/HPs datasets. In the case of *A. thaliana*, for example, two models (RPs/NRPs and RPs/HPs) were generated for each of the possible scenarios: *A. thaliana*/*A. thaliana* (10-fold cross validation); *A. thaliana*/*G. max*; *A. thaliana*/*S. lycopersicum*; *A. thaliana*/*O. sativa*;

A. thaliana/*Z. mays*; *A. thaliana*/*O. lucimarinus*. The same was done for each of the other species.

The measures of sensitivity (Equation 3), specificity (Equation 4), and F-measure (Equation 5) were also used for a broader assessment [15]. Those measures can be calculated using the amounts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The values collected are the weighted average of the positive class and the negative class.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Accuracy = \frac{TP + TN}{(TP + FN) + (FP + TN)} \tag{2}$$

$$Sensitivity\ or\ True\ Positive\ Rate\ (TPR) = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity\ or\ True\ Negative\ Rate\ (TNR) = \frac{TN}{FP + TN} \tag{4}$$

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

Ensemble Prediction Method

The previous section described the methodology by which 72 models were created with different algorithms and training datasets. Those models, along with the way they are employed, make up the Rama method. Rama comprises two steps, each performing a series of tasks, as illustrated in Figure 1.

Before carrying out steps 1 and 2, the models generated are loaded (Figure 1A and 1B). Step 1 is responsible for classifying a protein using the models trained with the RPs/NRPs datasets. In this step, the user informs a protein sequence and respective species (Figure 1C). Next, Rama selects the models that have been adjusted to the informed species for applying them on the input sequence (Figure 1D). Each of the models applied generates a probability of the amino acid sequence informed being a ribosomal protein. The generated probabilities are averaged to produce the final ensemble probability P. The sequence classification at this point will be determined according to the chosen discriminant probability (default value: 0.5). See Equation 6 and Figure 1E. If $P < 0.5$ (class 0), the program is finalized and classifies the amino acid sequence as a non-ribosomal protein (Figure 1F). However,

if $P \geq 0.5$ (class 1), the amino acid sequence is classified as a possible ribosomal protein (Figure 1G) and is given as input to step 2 (Figure 1H).

$$Class = \begin{cases} 1, & \frac{\sum_1^6 P_i}{6} \geq 0.5 \\ 0, & \frac{\sum_1^6 P_i}{6} < 0.5 \end{cases} \quad (6)$$

Step 2 applies ML models to classify the sequence coming from step 1 as ribosomal or histone. This is necessary because RPs have some characteristics in common with HPs. For instance, both are positively charged proteins, which interact with nucleic acids. RPs directly interact with RNAs that make up the ribosomes, whereas HPs form histone-DNA complexes for the chromatin assembly. As described for the first step, in step 2 the sequence is given to ML models adjusted to the species selected by the user, but trained with datasets made up of RPs/HPs (Figure 1I). Next, the probabilities generated are processed in the same way as in step 1 (Equation 6) to obtain the final classification (Figure 1J). Thus, if the class is 0, then the sequence is considered non-ribosomal (Figure 1L). On the other hand, if the class is 1, it means evidence that the amino acid sequence informed corresponds to a ribosomal protein; thereby, it is classified and displayed to the user as such.

The user can configure four parameters in Rama. The first is the selection of which models (species) will take part of the ensemble classification. The second parameter is the species to which the amino acid sequence informed by the user belongs. The third parameter allows setting the threshold value used in Equation 6. Finally, the fourth parameter defines whether the probabilities of the protein being a ribosomal protein according to each model will be displayed. Notice that there are models constructed for six plant species. However, the good results (presented further on) obtained in the inter-species experiments demonstrate that our computational tool might be used to make predictions for other species either using the built-in models.

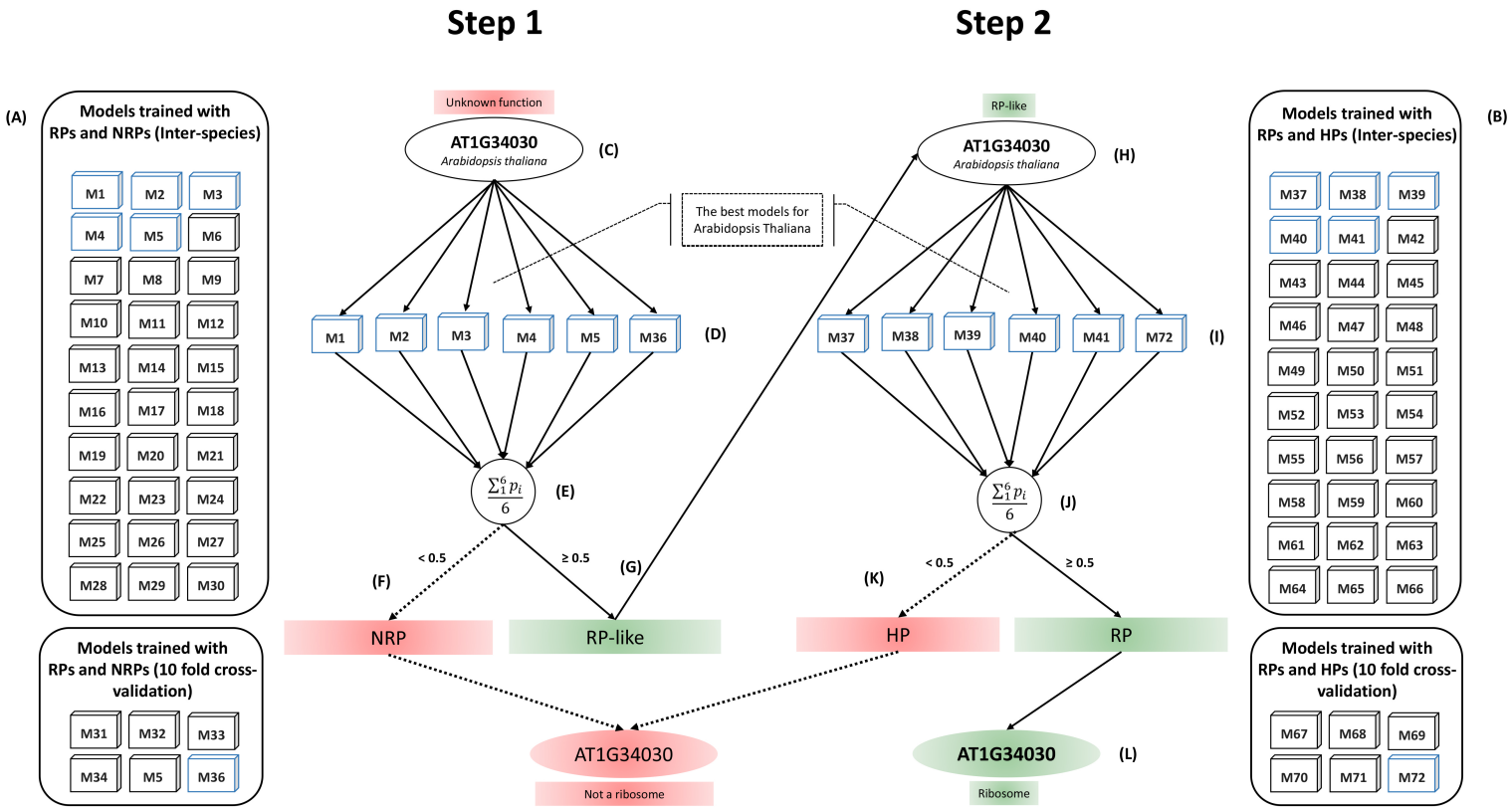

Figure 1 Illustration of the Rama method using a sequence of *A. thaliana* as an example. The proposed approach is described in 12 parts. In A and B, it is depicted the ML models built. Parts C, D, E, F, and G comprise Rama's step 1. In C, it is shown the input (amino acid sequence and species, e.g., *A. thaliana*) to the models for RPs/NRPs classification, according to the selected species. In D, the input protein is subject to these models. In E, the probabilities generated by the models are computed according to Equation 6, producing either class 0 or class 1. In case of class 0, i.e., not a ribosome protein (F), the program stops and informs the results. In case of class 1, i.e., a ribosome-like protein (G), the same amino acid sequence is used as input to the method's step 2, which is illustrated in parts H, I, J, K, and L. In H, the amino acid sequence is now given as input to the classification models for RPs/HPs (I). In J, the models' probabilities are aggregated to generate the final classification. It can be either 0, i.e., classified as a histone protein (K) or 1, i.e., classified as a ribosomal protein (L).

in vitro Nucleic Acid binding assay

Full-length candidate proteins AT3G51010, AT4G11385 and control proteins AT4G40030 (histone H3), AT1G29970 (RPL18) were fused to HA tag and expressed *in vitro* using the TnT in vitro transcription/ translation system (Promega). The DNA binding assay protocol was modified from [16]. Equal amounts of protein were incubated with either single-stranded or double-stranded deoxyribonucleic acid lyophilized powder attached to cellulose beads from calf thymus DNA (1 mg/ml). After incubation at 4 oC for 30 min, the beads were washed five times in RHPA buffer (10 mM Tris-HCl, pH 7.4, 2.5 mM MgCl2, 0.5% (v/v) Triton X-100) and then boiled in SDS loading buffer. The proteins were separated by SDS-PAGE detected by western blot using an anti-HA antibody. The RNA binding assay was performed according to [17]. Total RNA from *A. thaliana* were extracted and biotinylated using BrightStar Psolaren-Biotin kit, following the manufacture's instructions. For RNA–protein pulldown, biotinylated RNAs were first incubated with streptavidin-bound beads (Dynabeads; invitrogen) in IP100 buffer (100 mM potassium glutamate, 50 mM Tris-HCl pH 7.5, 100 mM NaCl, 0.2% (v/v) Nonidet P40) for 2 h at 4 oC, then washed five times in IP100 buffer. Equal amounts of protein were then added to RNA-bound beads and incubated under rotation at 4 oC for 30 min. Subsequently, beads were washed five times with IP100 buffer, boiled in SDS loading buffer and subjected to SDS-PAGE. The proteins were detected by western blot using an anti-HA antibody.

Results and Discussion

In Silico Results

To conduct an *in silico* evaluation of Rama, five computational experiments were performed: (i) Evaluation of the importance of each attribute using IG; (ii) inter-species tests and 10-fold cross validation of the models generated with the datasets composed of RPs/NRPs; (iii) inter-species tests and 10-fold cross validation of the models generated with the datasets composed of RPs/HPs; (iv) tests using the ensemble methodology proposed in Rama; and (v) use of Rama to check whether it would be capable of identifying proteins annotated as RPs in Phytozome v.11 using the models created from proteins of Phytozome v.10. Notice that such RPs from version 11, used as test set, had no function associated in version 10.

The first experiment allows examine the importance of the attributes for a better understanding of their influence in RP classification. These analyses were performed through the IG method. Table 3 presents the values obtained by the IG method for the RPs/NRPs datasets. Likewise, Table 4 describes the IG results for the RPs/HPs datasets. Besides IG values, Tables 3 and 4 display the position of each attribute in the ranking formed from the IG values so as to identify attributes with greatest influence on the classification process. When the ranking is analyzed, even though we can note differences among the analyzed species, the similarities related to top-ranked attributes are noteworthy. In most cases, the two best attributes in one species coincide with the two best attributes in another species. For example, Table 3 shows that the attributes positively_charged and amount_of_amino_acids obtained the best values in all species. In Table 4, those attributes were not unanimously the best ones, but they still stood out in general. The importance of the attribute

positively_charged, in particular, seems to corroborate the observation made previously about ribosome affinity to RNAs. The amount_of_amino_acids, in turn, also appears to strongly characterize the protein types.

Although there are attributes with zeroed IG value, removing them causes a slight negative variation in the performance of the models. This occurs because IG assesses the importance of an attribute independently, i.e., not in conjunction with other attributes. Therefore, the performance decline after removing one such attribute shows that the correlation of the used attributes is important for the classification task. Furthermore, no attribute presented IG equal to zero for all species under examination.

In the second experiment, we assessed the predictive power of the models created with RPs/NRPs. In this case, two validation tests were performed: 10-fold cross-validation and inter-species test, i.e., each set of proteins from one species was given as input to models trained from proteins of other species. This experiment allowed to measure the generality of the models that classify RPs/NRPs (first stage) using well-known measurements such as: Precision, accuracy, sensitivity, and specificity. As can be seen in Table 5, this generality is shown both for each individual model (values provided for each dataset used as training set) and as an overall evaluation by showing the mean values.

Analyzing the inter-species tests where *A. thaliana* was used as the test set, for instance, Table 5 shows the performance of each of the five models built from proteins of the other five species and also the overall performance of them (mean values). As a whole, these models achieved mean precision, accuracy, sensitivity, and specificity of 0.93, 0.94, 0.94, and 0.87, respectively. Still in Table 5, in the inter-species tests, we can observe that when performing the same type of experiment to identify RPs/NRPs of *O. sativa*, *O. lucimarinus*, *G. max*, and *S. lycopersicum* using models created from distinct species, we obtained average values greater than 0.90 for precision, accuracy as well as sensitivity, and values higher than 0.79 for specificity. Using the *Z. mays* proteins as test set, the values were slightly lower, resulting in mean precision, accuracy, sensitivity, and specificity of 0.87, 0.87, 0.87, and 0.75, respectively. The tests using 10-fold cross validations, also shown in Table 5, presented an overall performance of more than 0.92 for precision, accuracy as well as precision, and more than 0.83 for specificity.

The third experiment is analogous to the second one. This time, however, we assessed the generality of the models constructed from the datasets composed of RPs/HPs. Again, we show the results for the six best model setups obtained for each species, i.e., the model using the dataset of the species itself, evaluated by means of 10-fold cross validation, and the other five models from the five distinct species, using the proteins of the target species as test set. Table 6 shows, for instance, that the inter-species tests for *A. thaliana* have achieved mean precision, accuracy, sensitivity, and specificity of 0.90, 0.89, 0.89, and 0.72, respectively. Looking into the results of inter-species tests for *O. sativa*, *G. max*, and *S. lycopersicum*, the models adjusted for these species obtained average values greater than 0.90 of precision, accuracy, and sensitivity, and values greater than 0.69 of specificity. The results of the same type of test for *O. lucimarinus* and *Z. mays*, in turn, achieved an average classification performance somewhat reduced, reaching values greater than 0.85 of

precision, accuracy as well as sensitivity, and values greater than 0.55 of specificity (Table 6).

Table 6 shows also the cross-validation tests. As can be seen, the models obtained, on average, 0.94 of precision, 0.94 of accuracy, 0.94 of sensitivity, and 0.80 of specificity. These tests, similarly to the numbers shown in Table 5, led to better results, in general. This is expected because in cross-validation the proteins for training and test are from the same species.

The high values for sensitivity and precision ($\geq$ 0.85 and $\geq$0.86, on average, respectively) obtained from the models of RPs/NRPs and RPs/HPs (72 models presented in Table 5 and 6) make up an important feature of Rama. They mean that: i) Few RPs would be missed (high sensitivity), and ii) among the ones predicted as positives, few of them would be FPs, i.e., financial resources and time would be saved in laboratory validations. Nonetheless, Rama implements a combined use of these models to achieve even better results. As a fourth *in silico* experiment, we tested those machine learning models in an ensemble approach, as already described (Equation 6), i.e., running the complete pipeline implemented in Rama. Notice that Figure 1 (parts A and B) shows that the user may apply all six models optimized for a certain species, i.e., the model constructed using the proteins of the species itself, optimized by cross-validation tests, could (and should) be included. However, to provide evidence of robustness of the resultant models, our experiment did not include such a model for the species being evaluated. Table 7 shows the performance of each ensemble classification. The values in the first line, regarding *A. thaliana*, for instance, are a result of models created from *S. lycopersicum*, *Z. mays*, *O. lucimarinus*, *G. max*, and *O. sativa* (i.e., models M1, M2, M3, M4, M5, M37, M38, M39, M40, and M41 shown in Tables 5 and 6). These values represent the mean performance of the five models, averaging for positive and negative examples. As can be seen, using RPs and NRPs of *A. thaliana* as test set led to a very satisfactory performance: 0.94 of precision, accuracy as well as sensitivity, and 0.87 of specificity. For the other species, Rama could achieve values greater than 0.88 for precision, accuracy as well as sensitivity, and values greater than 0.75 for specificity. Therefore, sensitivity and precision could be further increased using the described ensemble approach. Even specificity could be improved.

In a fifth computational experiment, we used four proteins of *A. thaliana* with unknown function in Phytozome v.10, described as RPs in Phytozome v.11 as a test set for the Rama pipeline. Remember that the models built for this pipeline used proteins of Phytozome v.10 as training examples, i.e., the tested proteins of Phytozome v.11 did not make part of the learning process. These tested proteins were: AT3G02080.1, AT5G15520.1, AT5G61170.1, and ATCG00760.1. Rama correctly classified all of them as RPs. Thus, the proposed method was proven to be able to predict a change that occurred when Phytozome v.10 was updated to version 11.

Two novel predicted RPs by Rama displayed RNA binding activity consistent with the typical ribosomal function of RPs

Next, we searched for novel RPs by giving unknown protein sequences from *A. thaliana* to Rama as input, using the six models, optimized for *A. thaliana*. The

Additional file 3 displays the results of the top 50 ranked (by probability) protein sequences predicted as RPs. Notice that the results shown in Tables 5, 6 and 7 are from predictions using the default discriminant probability ($P \geq 0.5$). If the user wishes to obtain predictions with very high confidence, i.e., maximize precision, it is possible to choose a higher probability threshold. Due to limitation of time and funding, we used this possibility to randomly select two proteins among the ones with high probability ($P \geq 0.95$), AT3G51010 and AT4G11385, for further *in vitro* analyses. As fundamental components that make up and stabilize the structure of ribosomes, many ribosomal proteins contain characterized RNA binding domain, which is often surface-exposed domain interacting with multiple RNA elements within rRNA [18]. We designed an *in vitro* nucleic acid binding assay to further confirm that the new predicted RPs would share common biochemical properties of ribosomal proteins (Figure 2). We also included in the assay a Histone representative, H3, as a negative control and a known and well-characterized ribosomal protein, RPL10 (AT1G14320), as a positive control. To assay for DNA binding activity, the recombinant HA-fused proteins were transcribed and translated *in vitro* (input) and incubated with ssDNA or dsDNA linked to sepharose beads. RNA binding activity was monitored by incubating the *in vitro* translated proteins with *A. thaliana* biotinylated RNA conjugated with streptavidin beads. As expected, the histone H3 bound to dsDNA and ssDNA, but not to RNA. In contrast, AT1G14320 (the L10 positive control) and the new predicted RPs, AT3G51010 and AT4G11385, bound to RNA with high affinity, bound to ssDNA with low affinity, and did not interact with dsDNA. These results confirmed that the predicted new RPs display a profile of nucleic acid binding activity which is consistent with the role of ribosomal proteins. Furthermore, they demonstrated that the ribosomal proteins bind to ssDNA although with a much lower affinity than histones. These *in vitro* results show the need of the second stage of the pipeline implemented in Rama to distinguish histones and ribosomal proteins, since they share biochemical properties such as ssDNA binding activity. It is important to highlight that InterProScan could not predict proteins AT3G51010 and AT4G11385 as RPs.

Conclusions

This study presented a new method to predict RPs called Rama. This method uses several ML models to predict RPs in two distinct steps. In the first, RPs are distinguished from NRPs and, in the second, RPs are distinguished from HPs. The models used were trained and adjusted for five plant species and one phytoplankton species. The *in silico* experiments showed that Rama was able to differentiate RPs from other proteins for the six species with a high success rate. Besides the positive *in silico* results, Rama was able to successfully predict two ribosomal proteins, whose annotations were previously tagged as unknown function, that were experimentally confirmed *in vitro*. Interestingly, InterProScan could not identify the same proteins as RPs. It is also important to highlight that Rama is approximately 600 times faster than InterProScan. Considering the good results presented here using Rama, and that Rama's approach is quite different from the sequence homology procedure widely employed in the scientific community, we believe that the method hereby presented is promising and innovative, mainly as an important complement to traditional tools, because, additionally to the homology-based prediction performed by

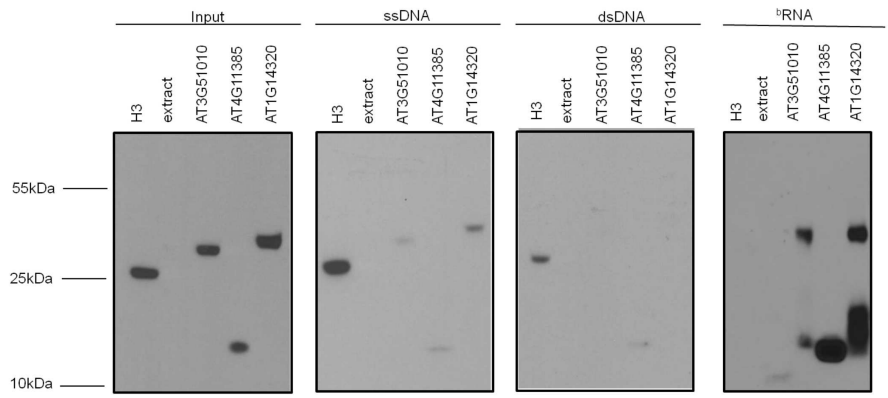

Figure 2 Nucleic acid binding activity of novel predicted RPs. HA fused-Histone H3, ribosomal protein RPL10 (AT1G14320), and the novel predicted RPs AT3G51010 and AT4G11385 (also fused to HA) were *in vitro* transcribed/translated (input), incubated with sepharose bound-ssDNA, sepahrose bound-dsDNA or *Arabidopsis* streptavidin-beads bound-biotinylated RNA (bRNA), pulled-down, separated by SDS-PAGE, and immunoblotted with anti-HA serum. Input shows an immunoblot of the *in vitro* transcribed/translated HA fusions. The sizes and positions of protein molecular mass markers are shown on the left.

such tools, Rama can identify completely novel proteins that are particular to the studied species. Even though our experiments included six species, the inter-species tests demonstrate that our models might be used to other species either. We believe that the good results obtained for one species using models from distinct species are a consequence of using classification attributes common to any protein of any species, making the resulting ML models sound predictors. No matter how robust and reliable tools such as InterProScan are, it is interesting to explore other options that may perform analyses far beyond the known methods.

Abbreviations

A. thaliana: *Arabidopsis thaliana* **G. max:** *Glycine max* **HMM:** hidden markov model **HPs:** histone proteins **IG:** information gain **ML:** machine Learning **MLP:** multilayer perceptron **mRNAs:** messenger RNA **NRPs:** not ribosomal proteins **O. sativa:** *Oryza sativa* **O. lucimarinus:** *Ostreococcus lucimarinus* **RF:** randon forest **RPs:** ribosomal proteins **rRNAs:** ribosomal RNA **SMO:** sequential minimal optimization **S. lycopersicum:** *Solanum lycopersicum* **SVM:** support vector machine **TPR:** true positive rate **TSs:** training sets **Z. mays:** *Zea mays*

Competing interests

The authors declare that they have no competing interests.

Author's contributions

JCFS suggested this study. JCFS, TFMC, FRC, and EPBF designed this investigation. JCFS and TFMC implemented the software and provided the *in silico* validation of the method. EPBF suggested and designed the *in vitro* experiment. IPC performed the *in vitro* experiment. All authors helped to draft the manuscript. FRC and EPBF supervised this study. All authors read and approved the final manuscript.

Acknowledgements

The authors acknowledge the support of the Universidade Federal de Viçosa (UFV), Instituto Nacional de Ciência e Tecnologia em Interações Planta-Praga (INCTIPP), Departamento de Informática – UFV, and Division of Support for Scientific and Technological Development of UFV (DCT). Acknowledgments to the financial support of Brazilian institutions: CAPES, CNPq, FAPEMIG, and INCTIPP

Author details

[1]Departamento de Informática, Universidade Federal de Viçosa, 36570-900, Viçosa, Brazil. [2]BIOAGRO/Instituto Nacional de Ciência e Tecnologia em Interação Planta-Praga, Universidade Federal de Viçosa, 36570-900, Viçosa, Brazil.

References

1. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J.D., Grimstone, A.: Molecular biology of the cell (3rd edn). Trends in Biochemical Sciences **20**(5), 210–210 (1995)
2. Rocha, C.S., Santos, A.A., Machado, J.P.B., Fontes, E.P.: The ribosomal protein l10/qm-like protein is a component of the nik-mediated antiviral signaling. Virology **380**(2), 165–169 (2008)
3. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., *et al.*: Interproscan 5: genome-scale protein function classification. Bioinformatics **30**(9), 1236–1240 (2014)
4. Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., *et al.*: The interpro protein families database: the classification resource after 15 years. Nucleic acids research **43**(D1), 213–221 (2015)
5. Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., *et al.*: Phytozome: a comparative platform for green plant genomics. Nucleic acids research **40**(D1), 1178–1186 (2012)
6. Tan, P.-N., *et al.*: Introduction to Data Mining. Pearson Education India, India (2006)
7. Platt, J., et al.: Sequential minimal optimization: A fast algorithm for training support vector machines (1998)
8. Nelson, D.L., Lehninger, A.L., Cox, M.M.: Lehninger Principles of Biochemistry. Macmillan, New York (2008)
9. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: ICML, vol. 97, pp. 412–420 (1997)
10. 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)
11. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
12. Kursa, M.B.: Robustness of random forest-based gene selection methods. BMC bioinformatics **15**(1), 1 (2014)
13. Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.-C.: psuc-lys: Predict lysine succinylation sites in proteins with pseaac and ensemble random forest approach. Journal of Theoretical Biology (2016)
14. 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)
15. Powers, D.M.: Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation (2011)
16. Kaiserli, E., Páldi, K., O'Donnell, L., Batalov, O., Pedmale, U.V., Nusinow, D.A., Kay, S.A., Chory, J.: Integration of light and photoperiodic signaling in transcriptional nuclear foci. Developmental cell **35**(3), 311–321 (2015)
17. Vert, G., Chory, J.: Downstream nuclear events in brassinosteroid signalling. Nature **441**(7089), 96–100 (2006)
18. Bandziulis, R., Swanson, M., Dreyfuss, G.: Rna-binding proteins as developmental regulators. Genes Dev **3**(4), 431–437 (1989)

Tables

Table 1 Training set composed of RPs and NRPs. Number of RPs and NRPs that make up the training set for each species.

Training set	Number of RPs	Number of NRPs
A. thaliana	516	1548
G. max	1085	3255
O. sativa	539	1617
O. lucimarinus	176	528
S. lycopersicum	485	1455
Z. mays	1408	4224

Table 2 Training set composed of RPs and HPs. Number of RPs and HPs that make up the training set for each species.

Training set	Number of RPs	Number of HPs
A. thaliana	516	182
G. max	1085	269
O. sativa	539	112
O. lucimarinus	176	46
S. lycopersicum	485	119
Z. mays	1408	395

Table 3 IG value for the datasets composed of RPs and NRPs. Values obtained when running the IG method in each training set. The rank created with IG values, shown in parentheses, highlights the importance of each attribute for each species.

Attribute	*A. thaliana*	*G. max*	*O. sativa*	*O. lucimarinus*	*S. lycopersicum*	*Z. mays*
aromatic	0.0949 (4°)	0.0509 (4°)	0.0307 (8°)	0.0249 (8°)	0.0622 (5°)	0.00987 (9°)
hydrophobic	0.0554 (7°)	0.0418 (7°)	0.0623 (5°)	0.0649 (5°)	0.0670 (4°)	0.05021 (5°)
molecular mass	0.0313 (8°)	0.0299 (8°)	0.0655 (4°)	0.0439 (6°)	0.0127 (8°)	0.06119 (4°)
negatively charged	0.0741 (5°)	0.0462 (6°)	0.0579 (6°)	0.1239 (3°)	0.0457 (6°)	0.02959 (7°)
nonpolar aliphatic	0.0000 (9°)	0.0000 (9°)	0.0210 (9°)	0.0260 (7°)	0.0000 (9°)	0.02596 (8°)
polar uncharged	0.1196 (3°)	0.1356 (3°)	0.0468 (7°)	0.0000 (9°)	0.0980 (3°)	0.03457 (6°)
positively charged	0.3450 (1°)	0.3123 (1°)	0.3284 (1°)	0.3107 (1°)	0.3076 (1°)	0.26796 (1°)
amount of amino acids	0.2832 (2°)	0.2365 (2°)	0.2501 (2°)	0.2245 (2°)	0.1909 (2°)	0.14375 (2°)
volume	0.0600 (6°)	0.0500 (5°)	0.0956 (3°)	0.0676 (4°)	0.0224 (7°)	0.08408 (3°)

Table 4 IG value for datasets composed of RPs and HPs. Values obtained when running the IG method in each training set. The rank created with IG values, shown in parentheses, highlights the importance of each attribute for each species.

Attribute	*A. thaliana*	*G. max*	*O. sativa*	*O. lucimarinus*	*S. lycopersicum*	*Z. mays*
aromatic	0.0000 (8°)	0.0733 (5°)	0.0000 (6°)	0.0000 (4°)	0.0357 (4°)	0.0431 (7°)
hydrophobic	0.0495 (6°)	0.0106 (9°)	0.0193 (5°)	0.0000 (4°)	0.0000 (6°)	0.0252 (9°)
molecular mass	0.0758 (5°)	0.0607 (7°)	0.0576 (4°)	0.0000 (4°)	0.0582 (3°)	0.0961 (3°)
negatively charged	0.1046 (4°)	0.0887 (4°)	0.0000 (6°)	0.0619 (3°)	0.0000 (6°)	0.0864 (4°)
nonpolar aliphatic	0.0000 (8°)	0.0398 (8°)	0.0000 (6°)	0.0000 (4°)	0.0000 (6°)	0.0315 (8°)
polar uncharged	0.0324 (7°)	0.0888 (3°)	0.0000 (6°)	0.0000 (4°)	0.0251 (5°)	0.0507 (6°)
positively charged	0.1218 (2°)	0.1261 (2°)	0.0945 (2°)	0.0930 (2°)	0.0000 (6°)	0.1560 (1°)
amount of amino acids	0.1094 (3°)	0.2398 (1°)	0.1279 (1°)	0.2072 (1°)	0.1559 (1°)	0.0526 (5°)
volume	0.1742 (1°)	0.0654 (6°)	0.0743 (3°)	0.0000 (4°)	0.0687 (2°)	0.1268 (2°)

Table 5 Results of the classification models for the RPs/NRPs datasets. Inter-species and 10-fold cross validation tests. The values are a weighted average of individual values for positive and negative examples.

	Model code	Trainning	Testing	Accuracy	Sensitivity	Precision	F-measure	Specificity
inter-species	M1	*Z. mays*	*A. thaliana*	0.9370	0.9370	0.9360	0.9370	0.8850
	M2	*O. sativa*	*A. thaliana*	0.9379	0.9380	0.9370	0.9370	0.8670
	M3	*O. lucimarinus*	*A. thaliana*	0.9437	0.9440	0.9430	0.9430	0.8900
	M4	*G. max*	*A. thaliana*	0.9404	0.9400	0.9400	0.9390	0.8650
	M5	*S. lycopersicum*	*A. thaliana*	0.9433	0.9430	0.9430	0.9420	0.8670
	Average			0.9405	0.9404	0.9398	0.9396	0.8748
	M6	*A. thaliana*	*Z. mays*	0.8785	0.8790	0.8760	0.8760	0.7650
	M7	*O. sativa*	*Z. mays*	0.8865	0.8870	0.8850	0.8810	0.7260
	M8	*O. lucimarinus*	*Z. mays*	0.8790	0.8790	0.8770	0.8770	0.7730
	M9	*G. max*	*Z. mays*	0.8565	0.8570	0.8560	0.8560	0.7590
	M10	*S. lycopersicum*	*Z. mays*	0.8703	0.8700	0.8670	0.8680	0.7510
	Average			0.8742	0.8744	0.8722	0.8716	0.7548
	M11	*A. thaliana*	*O. sativa*	0.9239	0.9240	0.9230	0.9230	0.8500
	M12	*Z. mays*	*O. sativa*	0.9308	0.9310	0.9310	0.9290	0.8330
	M13	*O. lucimarinus*	*O. sativa*	0.9179	0.9180	0.9160	0.9160	0.8280
	M14	*G. max*	*O. sativa*	0.9174	0.9170	0.9190	0.9180	0.8780
	M15	*S. lycopersicum*	*O. sativa*	0.9128	0.9130	0.9120	0.9120	0.8410
	Average			0.9205	0.9206	0.9202	0.9196	0.8460
	M16	*A. thaliana*	*O. lucimarinus*	0.9247	0.9250	0.9240	0.9240	0.8650
	M17	*Z. mays*	*O. lucimarinus*	0.9318	0.9320	0.9310	0.9300	0.8480
	M18	*O. sativa*	*O. lucimarinus*	0.9303	0.9300	0.9300	0.9280	0.8290
	M19	*G. max*	*O. lucimarinus*	0.9176	0.9180	0.9160	0.9170	0.8440
	M20	*S. lycopersicum*	*O. lucimarinus*	0.9176	0.9180	0.9160	0.9160	0.8290
	Average			0.9244	0.9246	0.9234	0.9230	0.8430
	M21	*A. thaliana*	*G. max*	0.9156	0.9160	0.9150	0.9130	0.8080
	M22	*Z. mays*	*G. max*	0.9131	0.9130	0.9130	0.9100	0.7910
	M23	*O. sativa*	*G. max*	0.9147	0.9150	0.9140	0.9120	0.8030
	M24	*O. lucimarinus*	*G. max*	0.9133	0.9130	0.9120	0.9120	0.8240
	M25	*S. lycopersicum*	*G. max*	0.9223	0.9220	0.9210	0.9200	0.8270
	Average			0.9158	0.9158	0.9150	0.9134	0.8106
	M26	*A. thaliana*	*S. lycopersicum*	0.9154	0.9150	0.9140	0.9130	0.8070
	M27	*Z. mays*	*S. lycopersicum*	0.9015	0.9020	0.9010	0.8970	0.7610
	M28	*O. sativa*	*S. lycopersicum*	0.9087	0.9090	0.9070	0.9060	0.7980
	M29	*O. lucimarinus*	*S. lycopersicum*	0.9067	0.9070	0.9050	0.9040	0.7890
	M30	*G. max*	*S. lycopersicum*	0.9175	0.9180	0.9170	0.9150	0.8030
	Average			0.9100	0.9102	0.9088	0.9070	0.7916
Cross-Validation	M31	*Z. mays*		0.9090	0.9090	0.9070	0.9070	0.7980
	M32	*O. sativa*		0.9304	0.9300	0.9300	0.9290	0.8480
	M33	*O. lucimarinus*		0.9261	0.9260	0.9250	0.9240	0.8310
	M34	*G. max*		0.9476	0.9480	0.9470	0.9470	0.8900
	M35	*S. lycopersicum*		0.9092	0.9090	0.9080	0.9060	0.7920
	M36	*A. thaliana*		0.9520	0.9520	0.9530	0.9510	0.8750
	Average			0.9291	0.9290	0.9283	0.9273	0.8390

Table 6 Results of the classification models created for the RPs/HPs datasets. Inter-species and 10-fold cross validation tests. The values are a weighted average of individual values for positive and negative examples.

	Model code	Trainning	Testing	Accuracy	Sensitivity	Precision	F-measure	Specificity
inter-species	M37	*Z. mays*	*A. thaliana*	0.9312	0.9310	0.9340	0.9280	0.8230
	M38	*O. sativa*	*A. thaliana*	0.8896	0.8900	0.8970	0.8810	0.7050
	M39	*O. lucimarinus*	*A. thaliana*	0.8653	0.8650	0.8820	0.8490	0.6250
	M40	*G. max*	*A. thaliana*	0.9140	0.9140	0.9170	0.9100	0.7810
	M41	*S. lycopersicum*	*A. thaliana*	0.8796	0.8800	0.8880	0.8690	0.6800
	Average			0.8959	0.8960	0.9036	0.8874	0.7228
	M42	*A. thaliana*	*Z. mays*	0.8824	0.8820	0.8800	0.8810	0.7520
	M43	*O. sativa*	*Z. mays*	0.8707	0.8710	0.8790	0.8520	0.5610
	M44	*O. lucimarinus*	*Z. mays*	0.8003	0.8000	0.8410	0.7280	0.2880
	M45	*G. max*	*Z. mays*	0.8790	0.8790	0.8770	0.8680	0.6310
	M46	*S. lycopersicum*	*Z. mays*	0.8663	0.8660	0.8720	0.8470	0.5530
	Average			0.8774	0.8596	0.8698	0.8352	0.5570
	M47	*A. thaliana*	*O. sativa*	0.9308	0.9310	0.9330	0.9320	0.8580
	M48	*Z. mays*	*O. sativa*	0.9431	0.9430	0.9420	0.9420	0.8260
	M49	*O. lucimarinus*	*O. sativa*	0.8940	0.8940	0.8880	0.8840	0.5890
	M50	*G. max*	*O. sativa*	0.9462	0.9460	0.9450	0.9440	0.7910
	M51	*S. lycopersicum*	*O. sativa*	0.9216	0.9220	0.9220	0.9140	0.6580
	Average			0.9271	0.9272	0.9260	0.9232	0.7444
	M52	*A. thaliana*	*O. lucimarinus*	0.9099	0.9100	0.9070	0.9080	0.7680
	M53	*Z. mays*	*O. lucimarinus*	0.9099	0.9100	0.9150	0.9010	0.6710
	M54	*O. sativa*	*O. lucimarinus*	0.9054	0.9050	0.9150	0.8940	0.6380
	M55	*G. max*	*O. lucimarinus*	0.8963	0.8960	0.9080	0.8820	0.6040
	M56	*S. lycopersicum*	*O. lucimarinus*	0.8738	0.8740	0.8840	0.8530	0.5330
	Average			0.8990	0.8990	0.9058	0.8876	0.6428
	M57	*Z. mays*	*G. max*	0.9364	0.9360	0.9360	0.9340	0.7940
	M58	*O. sativa*	*G. max*	0.9194	0.9190	0.9220	0.9130	0.7000
	M59	*O. lucimarinus*	*G. max*	0.8943	0.8940	0.8900	0.8870	0.6580
	M60	*S. lycopersicum*	*G. max*	0.9313	0.9310	0.9350	0.9260	0.7340
	M61	*A. thaliana*	*G. max*	0.9194	0.9190	0.9200	0.9200	0.8320
	Average			0.9202	0.9198	0.9206	0.9160	0.7436
	M62	*A. thaliana*	*S. lycopersicum*	0.9006	0.9010	0.8970	0.8980	0.7350
	M63	*Z. mays*	*S. lycopersicum*	0.9188	0.9190	0.9210	0.9120	0.6950
	M64	*O. sativa*	*S. lycopersicum*	0.9238	0.9240	0.9280	0.9170	0.7020
	M65	*O. lucimarinus*	*S. lycopersicum*	0.8841	0.8840	0.8850	0.8690	0.5720
	M66	*G. max*	*S. lycopersicum*	0.9321	0.9320	0.9330	0.9280	0.7550
	Average			0.9119	0.9120	0.9128	0.9048	0.6918
Cross-Validation	M67	*Z. mays*		0.9251	0.9250	0.9260	0.9210	0.7640
	M68	*O. sativa*		0.9539	0.9540	0.9540	0.9520	0.7990
	M69	*O. lucimarinus*		0.9189	0.9190	0.9170	0.9170	0.7860
	M70	*G. max*		0.9667	0.9670	0.9670	0.9660	0.8740
	M71	*S. lycopersicum*		0.9188	0.9190	0.9230	0.9110	0.6820
	M72	*A. thaliana*		0.9570	0.9570	0.9570	0.9560	0.8990
	Average			0.9401	0.9402	0.9407	0.9372	0.8007

Table 7 Results obtained by applying the whole pipeline (stage 1 and 2) implemented in Rama. The column 'Species' describes the target species, i.e., the one whose proteins were used as test set for the models created from the other five species. The values are the mean measures of the five distinct models for each target species, averaging for positive and negative examples.

Species	Accuracy	Sensitivity	Precision	F-measure	Specificity
A. thaliana	0.9428	0.9428	0.9423	0.9418	0.8737
G. max	0.9195	0.9196	0.9190	0.9169	0.8091
O. lucimarinus	0.9332	0.9332	0.9323	0.9322	0.8603
O. sativa	0.9350	0.9351	0.9344	0.9337	0.8534
S. lycopersicum	0.9128	0.9129	0.9122	0.9097	0.7923
Z. mays	0.8849	0.8849	0.8817	0.8813	0.7528

Additional Files

Additional file 1 — Algorithms and parameters used to create thousands of models.

Additional file 2 — Algorithms and parameters used to create the 72 models of Rama, i.e., the best models obtained from tests described in Additional file 1.

Additional file 3 — Results of the top 50 ranked protein sequences of *A. thaliana* with unknown function submitted to Rama which were predicted as RP.

Capítulo 5

Conclusões

5.1 Conclusões gerais

A busca de novos conhecimentos ou informações confiáveis é uma tarefa constante e por várias vezes árdua. Neste sentido, as técnicas de DM são de suma importância e proporcionam encontrar informações desconhecidas que seriam impossíveis de serem encontradas de forma manual.

Nos três trabalhos apresentados nesta dissertação, encontram-se problemas práticos que de alguma forma podem prejudicar diversos estudos. Por exemplo, no primeiro trabalho apresentado, mostrou-se que muitas informações referentes a geminivírus estão despadronizadas e possuem inconsistência. Tal problema pode interferir diretamente em estudos baseados em homologia de sequência. Já no segundo trabalho, o qual utiliza NLP, foi possível ver que uma quantidade muito grande de artigos em uma área (como geminivírus) pode dificultar o pesquisador a perceber como determinado assunto está sendo retratado na área. Por fim, no trabalho do Rama, verificou-se a importância das proteínas ribossomais e que, atualmente, não existem técnicas específicas para identificá-las. Sendo assim, os métodos existentes nem sempre são capazes de identificar novas proteínas ribossomais utilizando apenas homologia de sequência.

Para tentar solucionar cada um desses problemas, criaram-se ferramentas online que foram disponibilizadas de forma livre à comunidade científica. Em cada uma destas ferramentas, foram empregadas técnicas de DM para corrigir e desco-

brir novas informações. Com a ferramenta intitulada *geminivirus.org* foi possível realizar uma série de procedimentos que permitem localizar e corrigir diversas informações que foram anotadas erradas. Tais informações são manipuladas utilizando conjuntos de regras com técnicas clássicas de DM. Ao final de todas as pesquisas e codificações realizadas para desenvolver o *geminivirus.org*, criou-se um banco de dados robusto e confiável onde os pesquisadores têm acesso a informações refinadas e um conjunto de ferramentas que auxiliam a pesquisa. Vale ressaltar que esse *Data Warehouse* estará em constante evolução, sendo pela adição de novas sequências de vírus e adição de novas funcionalidades.

Já a ferramenta criada utilizando NLP para relacionar palavras de resumos de artigos fornece ao pesquisador uma forma interativa e clara de analisar se uma determinada linha de pesquisa está sendo tratada na área em questão. Também foi visto que é interessante relacionar duas áreas de estudos de patógenos diferentes para observar como um determinado assunto é tratado em cada área ou até mesmo encontrar relacionamento entre elas.

Também utilizando técnicas de DM, porém com foco na ML, desenvolveu-se o trabalho que deu origem ao Rama. Neste trabalho, conseguiu-se identificar RPs com alta taxa de sucesso utilizando um conjunto de modelos criados a partir de métodos de ML treinados com características fundamentais de qualquer proteína. A metodologia com que os modelos de ML foram dispostos proporcionou a criação de uma nova abordagem de classificadores em conjunto (*ensemble*) que permitiu alcançar resultados animadores. Tal metodologia se ajustou tão bem para predição de RPs que possivelmente também se ajustará para outros tipos de proteínas.

Todos esses trabalhos mostraram a capacidade e benefícios de utilizar técnicas de DM, além de demostrar que elas devem ser exploradas de forma inovadora e criativa no meio biológico. Observou-se também que técnicas de aprendizado de máquina podem contribuir significativamente para encontrar funções de proteínas desconhecidas, como comprovado nesse trabalho para RPs.

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.

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

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

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.

Ma, C.; Zhang, H. H. & Wang, X. (2014). Machine learning for big data analytics in plants. *Trends in plant science*, 19(12):798--808.

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

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

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.

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.

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 seguir são apresentados os arquivos suplementares do artigo **Rama: A machine learning approach for ribosomal protein prediction in plants.**

A.1 Additional file 1

Algorithms and parameters used to create thousands of models.

Algorithm	Parameter	Parameter	Variation	Minimum value	Maximum value
Sequential Minimal Optimization	Kernel: "PolyKernel"/ "NormalizedPolyKernel"/ "Puk"	C	x10	0.1	5000.0
			x10	0.5	5000.0
		Cache Size	+50000	0.0	400000.0
	Kernel: "RBFKernel"	Exponent	+0.5	0.1	4.0
		C	x10	0.1	5000.0
			x10	0.5	5000.0
Multilayer Perceptron	hiddenLayers: "a"/ "t"/ "i"/ "o"	Cache Size	+50000	0.0	400000.0
		Gamma	(+0.5) / 10	0.1	4.0
		Learning Rate	+0.02	0.01	1.0
		Momentum	+0.1	0.1	1.0
		Training Time	+500	500.0	1500.0
Randon Forest		Num Trees	+10	1.0	601.0
		Num Features	+1	0.0	10.0
		Seed	+10	1.0	601.0

A.2 Additional file 2

Algorithms and parameters used to create the 72 models of Rama, i.e., the best models obtained from tests described in Additional file 1.

Data set	Code	Trainning	Testing	Method	Weka parameters
Rps / NRPs	M1	Zmays	Athaliana	MLP	-L 0.15 -M 0.1 -N 1000 -E 20 -H t
Rps / NRPs	M2	Osativa	Athaliana	MLP	-L 0.1699 -M 0.6 -N 500 -E 20 -H t
Rps / NRPs	M3	Olucimarinus	Athaliana	MLP	-L 0.1699 -M 0.2 -N 500 -E 20 -H i
Rps / NRPs	M4	Gmax	Athaliana	RF	-I 31 -K 6 -S 101
Rps / NRPs	M5	Slycopersicum	Athaliana	RF	-I 61 -K 1 -S 481
Rps / NRPs	M6	Athaliana	Zmays	SMO	-C 5000.0 -K "NormalizedPolyKernel -C 0 -E 2.1"
Rps / NRPs	M7	Osativa	Zmays	SMO	-C 10.0 -K "Puk -C 0 -O 1.0 -S 0.1"
Rps / NRPs	M8	Olucimarinus	Zmays	MLP	-L 0.13 -M 0.6 -N 1000 -E 20 -H a
Rps / NRPs	M9	Gmax	Zmays	MLP	-L 0.5700 -M 0.7999 -N 500 -E 20 -H t
Rps / NRPs	M10	Slycopersicum	Zmays	RF	-I 251 -K 1 -S 491
Rps / NRPs	M11	Athaliana	Osativa	SMO	-C 50.0 -K "Puk -C 350000 -O 1.0 -S 2.1"
Rps / NRPs	M12	Zmays	Osativa	RF	-I 41 -K 3 -S 361
Rps / NRPs	M13	Olucimarinus	Osativa	MLP	-L 0.2299 -M 0.1 -N 1000 -E 20 -H t
Rps / NRPs	M14	Gmax	Osativa	RF	-I 71 -K 2 -S 171
Rps / NRPs	M15	Slycopersicum	Osativa	SMO	-C 5000.0 -K "PolyKernel -C 0 -E 2.1"
Rps / NRPs	M16	Athaliana	Olucimarinus	SMO	-C 50.0 -K "Puk -C 0 -O 1.0 -S 2.1"
Rps / NRPs	M17	Zmays	Olucimarinus	MLP	-L 0.9300 -M 0.6 -N 1500 -E 20 -H a
Rps / NRPs	M18	Osativa	Olucimarinus	MLP	-L 0.6900 -M 0.5 -N 500 -E 20 -H a
Rps / NRPs	M19	Gmax	Olucimarinus	MLP	-L 0.7700 -M 0.6 -N 1000 -E 20 -H o
Rps / NRPs	M20	Slycopersicum	Olucimarinus	MLP	-L 0.7500 -M 0.2 -N 1500 -E 20 -H i
Rps / NRPs	M21	Athaliana	Gmax	RF	-I 81 -K 2 -S 91
Rps / NRPs	M22	Zmays	Gmax	MLP	-L 0.4300 -M 0.8999 -N 500 -E 20 -H t
Rps / NRPs	M23	Osativa	Gmax	SMO	-C 10.0 -K "Puk -C 0 -O 1.0 -S 1.1"
Rps / NRPs	M24	Olucimarinus	Gmax	MLP	-L 0.05 -M 0.7 -N 500 -E 20 -H i
Rps / NRPs	M25	Slycopersicum	Gmax	RF	-I 161 -K 6 -S 11
Rps / NRPs	M26	Athaliana	Slycopersicum	RF	-I 111 -K 1 -S 91

Rps / NRPs	M27	Zmays	Slycopersicum	RF	-I 51 -K 1 -S 1
Rps / NRPs	M28	Osativa	Slycopersicum	MLP	-L 0.4700 -M 0.3000 -N 500 -E 20 -H i
Rps / NRPs	M29	Olucimarinus	Slycopersicum	MLP	-L 0.9700 -M 0.3000 -N 500 -E 20 -H t
Rps / NRPs	M30	Gmax	Slycopersicum	RF	-I 61 -K 1 -S 551
Rps / NRPs	M31	Zmays	-	SMO	-C 10.0 -K "Puk -C 350000 -O 1.0 -S 0.1"
Rps / NRPs	M32	Osativa	-	RF	-I 121 -K 2 -S 61
Rps / NRPs	M33	Olucimarinus	-	RF	-I 151 -K 2 -S 541
Rps / NRPs	M34	Gmax	-	SMO	-C 10.0 -K "Puk -C 250000 -O 1.0 -S 0.1"
Rps / NRPs	M35	Slycopersicum	-	RF	-I 181 -K 2 -S 431
Rps / NRPs	M36	Athaliana	-	SMO	-C 10.0 -K "Puk -C 0 -O 1.0 -S 0.1"
RPS / Hps	M37	Zmays	Athaliana	RF	-I 81 -K 10 -S 521
RPS / Hps	M38	Osativa	Athaliana	RF	-I 31 -K 9 -S 191
RPS / Hps	M39	Olucimarinus	Athaliana	RF	-I 31 -K 1 -S 321
RPS / Hps	M40	Gmax	Athaliana	RF	-I 11 -K 7 -S 71
RPS / Hps	M41	Slycopersicum	Athaliana	RF	-I 11 -K 9 -S 361
RPS / Hps	M42	Athaliana	Zmays	RF	-I 31 -K 5 -S 351
RPS / Hps	M43	Osativa	Zmays	RF	-I 41 -K 1 -S 91
RPS / Hps	M44	Olucimarinus	Zmays	SMO	-C 10.0 -K "Puk -C 0 -O 1.0 -S 0.1"
RPS / Hps	M45	Gmax	Zmays	RF	-I 31 -K 1 -S 281
RPS / Hps	M46	Slycopersicum	Zmays	RF	-I 91 -K 1 -S 391
RPS / Hps	M47	Athaliana	Osativa	RF	-I 131 -K 3 -S 321
RPS / Hps	M48	Zmays	Osativa	RF	-I 21 -K 2 -S 231
RPS / Hps	M49	Olucimarinus	Osativa	RF	-I 71 -K 2 -S 311
RPS / Hps	M50	Gmax	Osativa	RF	-I 21 -K 2 -S 171
RPS / Hps	M51	Slycopersicum	Osativa	RF	-I 491 -K 2 -S 441
RPS / Hps	M52	Athaliana	Olucimarinus	RF	-I 31 -K 2 -S 171
RPS / Hps	M53	Zmays	Olucimarinus	RF	-I 11 -K 1 -S 421
RPS / Hps	M54	Osativa	Olucimarinus	RF	-I 21 -K 6 -S 391
RPS / Hps	M55	Gmax	Olucimarinus	RF	-I 51 -K 2 -S 541
RPS / Hps	M56	Slycopersicum	Olucimarinus	RF	-I 101 -K 1 -S 371
RPS / Hps	M57	Zmays	Gmax	SMO	-C 5.0 -K "Puk -C 0 -O 1.0 -S 0.6"
RPS / Hps	M58	Osativa	Gmax	RF	-I 11 -K 5 -S 291
RPS / Hps	M59	Olucimarinus	Gmax	RF	-I 11 -K 1 -S 381
RPS / Hps	M60	Slycopersicum	Gmax	RF	-I 51 -K 3 -S 151
RPS / Hps	M61	Athaliana	Gmax	RF	-I 11 -K 6 -S 291

RPS / Hps	M62	Athaliana	Slycopersicum	RF	-I 21 -K 0 -S 161
RPS / Hps	M63	Zmays	Slycopersicum	RF	-I 31 -K 4 -S 21
RPS / Hps	M64	Osativa	Slycopersicum	RF	-I 81 -K 6 -S 481
RPS / Hps	M65	Olucimarinus	Slycopersicum	RF	-I 11 -K 1 -S 381
RPS / Hps	M66	Gmax	Slycopersicum	RF	-I 21 -K 4 -S 591
RPS / Hps	M67	Zmays	-	RF	-I 121 -K 1 -S 601
RPS / Hps	M68	Osativa	-	RF	-I 41 -K 2 -S 221
RPS / Hps	M69	Olucimarinus	-	RF	-I 31 -K 7 -S 171
RPS / Hps	M70	Gmax	-	RF	-I 81 -K 2 -S 61
RPS / Hps	M71	Slycopersicum	-	RF	-I 21 -K 1 -S 481
RPS / Hps	M72	Athaliana	-	RF	-I 31 -K 5 -S 401

A.3 Additional file 3

Results of the top 50 ranked protein sequences of *A. thaliana* with unknown function submitted to Rama which were predicted as RP

Columns P1 and P2 show the probabilities produced in stage 1 and 2, respectively, i.e., by RPs/NRPs and RPs/HPs models. The rows highlighted in yellow refers the new ribosome proteins validated in the in vitro experiment.

ID	Athaliana	P1	P2	Olucimarinus	P1	P2	Osativa	P1	P2	Slycopersicum	P1	P2	Gmax	P1	P2	Zmays	P1	P2	Classification	P1	P2
AT1G12938.1	RP	1.00	1.00	RP	1.00	1.00	RP	1.00	1.00	RP	0.93	0.91	RP	1.00	1.00	RP	0.99	0.98	RP	0.99	0.98
AT2G41231.1	RP	1.00	1.00	RP	1.00	1.00	RP	0.99	1.00	RP	1.00	0.91	RP	1.00	1.00	RP	0.98	1.00	RP	0.99	0.98
AT3G25716.1	RP	1.00	1.00	RP	1.00	1.00	RP	1.00	1.00	RP	0.97	0.91	RP	1.00	1.00	RP	0.98	1.00	RP	0.99	0.98
AT2G47485.1	RP	1.00	1.00	RP	1.00	0.87	RP	1.00	0.94	RP	0.95	1.00	RP	1.00	1.00	RP	1.00	1.00	RP	0.99	0.97
AT3G19660.1	RP	1.00	0.94	RP	1.00	0.97	RP	1.00	1.00	RP	0.95	0.91	RP	1.00	0.91	RP	0.99	0.99	RP	0.99	0.95
AT3G51010.1	RP	1.00	1.00	RP	1.00	0.97	RP	1.00	0.92	RP	0.93	0.91	RP	1.00	0.91	RP	0.98	0.95	RP	0.99	0.94
AT5G48860.1	RP	1.00	1.00	RP	1.00	0.97	RP	1.00	0.94	RP	0.92	0.91	RP	0.97	1.00	RP	0.98	1.00	RP	0.98	0.97
AT2G34800.1	RP	1.00	0.84	RP	1.00	1.00	RP	1.00	1.00	RP	0.95	1.00	RP	0.97	1.00	RP	0.99	1.00	RP	0.98	0.97
AT2G15930.1	RP	1.00	0.97	RP	1.00	0.94	RP	1.00	0.96	RP	0.93	1.00	RP	0.97	0.91	RP	0.97	1.00	RP	0.98	0.96
AT1G70270.1	RP	1.00	1.00	RP	1.00	0.9	RP	1.00	0.93	RP	0.95	0.91	RP	0.97	0.91	RP	0.98	0.99	RP	0.98	0.94
AT2G45126.1	RP	1.00	0.97	RP	1.00	1.00	RP	0.99	1.00	RP	0.89	0.91	RP	1.00	0.82	RP	0.99	0.91	RP	0.98	0.93
AT1G73940.1	RP	1.00	0.97	RP	1.00	0.87	RP	1.00	0.92	RP	0.92	0.82	RP	0.97	1.00	RP	0.98	0.9	RP	0.98	0.91
AT5G49410.1	RP	1.00	1.00	RP	1.00	0.97	RP	1.00	0.98	RP	0.93	1.00	RP	0.9	1.00	RP	0.97	1.00	RP	0.97	0.99
AT2G22241.1	RP	1.00	0.94	RP	1.00	1.00	RP	1.00	1.00	RP	0.92	1.00	RP	1.00	1.00	RP	0.93	1.00	RP	0.97	0.99
AT5G01881.1	RP	1.00	0.97	RP	0.99	0.9	RP	1.00	0.99	RP	0.85	1.00	RP	1.00	1.00	RP	0.98	1.00	RP	0.97	0.98
AT3G50764.1	RP	1.00	0.94	RP	1.00	0.97	RP	1.00	0.99	RP	0.82	1.00	RP	1.00	0.91	RP	0.99	0.99	RP	0.97	0.97
AT4G09860.1	RP	1.00	1.00	RP	1.00	1.00	RP	1.00	0.98	RP	0.92	0.91	RP	0.9	0.91	RP	0.99	1.00	RP	0.97	0.97

AT3G01240.1	RP	1.00	0.94	RP	1.00	0.97	RP	0.99	0.94	RP	0.85	0.91	RP	1.00	1.00	RP	1.00	0.95	RP	0.97	0.95
AT4G22510.1	RP	1.00	0.94	RP	0.97	1.00	RP	0.95	1.00	RP	0.92	0.91	RP	1.00	0.82	RP	0.96	1.00	RP	0.97	0.94
AT2G43060.1	RP	1.00	0.94	RP	0.99	0.94	RP	0.96	0.87	RP	0.92	1.00	RP	0.97	1.00	RP	0.96	0.93	RP	0.97	0.94
AT2G19820.1	RP	1.00	1.00	RP	1.00	1.00	RP	0.99	1.00	RP	0.84	1.00	RP	0.97	1.00	RP	0.99	1.00	RP	0.96	1.00
AT2G35208.1	RP	1.00	1.00	RP	1.00	0.94	RP	1.00	0.99	RP	0.87	0.91	RP	0.94	1.00	RP	0.97	1.00	RP	0.96	0.97
AT2G16015.1	RP	1.00	1.00	RP	1.00	0.87	RP	0.99	0.9	RP	0.79	1.00	RP	1.00	1.00	RP	0.99	1.00	RP	0.96	0.96
AT1G68845.1	RP	1.00	0.94	RP	1.00	0.94	RP	1.00	0.99	RP	0.8	1.00	RP	1.00	0.91	RP	0.98	1.00	RP	0.96	0.96
AT3G02242.1	RP	1.00	0.97	RP	1.00	0.97	RP	1.00	0.91	RP	0.8	0.82	RP	0.9	1.00	RP	0.99	0.91	RP	0.95	0.93
AT2G07776.2	RP	1.00	1.00	RP	1.00	0.9	RP	1.00	0.99	RP	0.75	1.00	RP	0.94	1.00	RP	0.97	1.00	RP	0.94	0.98
AT4G11385.1	RP	1.00	1.00	RP	1.00	0.9	RP	1.00	0.98	RP	0.74	1.00	RP	0.94	1.00	RP	0.99	1.00	RP	0.94	0.98
ATMG00530.1	RP	1.00	1.00	RP	1.00	0.9	RP	1.00	0.99	RP	0.75	1.00	RP	0.94	1.00	RP	0.97	1.00	RP	0.94	0.98
AT5G14410.1	RP	1.00	0.94	RP	1.00	0.87	RP	1.00	1.00	RP	0.77	1.00	RP	0.87	1.00	RP	0.98	1.00	RP	0.94	0.97
AT1G20875.1	RP	1.00	0.94	RP	1.00	1.00	RP	1.00	1.00	RP	0.79	1.00	RP	0.9	1.00	RP	0.92	0.99	RP	0.93	0.99
AT3G05936.1	RP	1.00	1.00	RP	1.00	0.97	RP	1.00	1.00	RP	0.8	1.00	RP	0.81	1.00	RP	0.97	1.00	RP	0.93	0.99
AT5G51960.1	RP	1.00	1.00	RP	0.96	0.97	RP	0.95	1.00	RP	0.8	1.00	RP	0.84	1.00	RP	0.99	0.99	RP	0.92	0.99
AT1G45165.1	RP	1.00	1.00	RP	1.00	0.97	RP	1.00	1.00	RP	0.72	1.00	RP	0.81	1.00	RP	0.98	1.00	RP	0.92	0.99
AT5G57760.1	RP	1.00	1.00	RP	1.00	0.84	RP	0.95	0.98	RP	0.67	1.00	RP	0.9	1.00	RP	0.99	1.00	RP	0.92	0.97
AT1G50691.1	RP	1.00	1.00	RP	1.00	0.97	RP	1.00	0.8	RP	0.85	1.00	RP	0.71	1.00	RP	0.97	0.96	RP	0.92	0.96
AT1G78476.1	RP	1.00	0.81	RP	0.99	1.00	RP	1.00	1.00	RP	0.79	0.91	RP	0.81	1.00	RP	0.95	1.00	RP	0.92	0.95
AT3G43291.1	RP	1.00	0.81	RP	0.99	0.84	RP	1.00	0.88	RP	0.82	0.73	RP	0.77	1.00	RP	0.95	0.95	RP	0.92	0.87
AT1G32260.1	RP	1.00	1.00	RP	0.98	0.97	RP	1.00	1.00	RP	0.82	1.00	RP	0.71	1.00	RP	0.96	1.00	RP	0.91	0.99
AT4G08351.1	RP	1.00	1.00	RP	1.00	0.74	RP	1.00	0.97	RP	0.67	1.00	RP	0.81	1.00	RP	1.00	0.98	RP	0.91	0.95
AT3G24929.2	RP	1.00	1.00	RP	0.99	0.94	RP	1.00	1.00	RP	0.69	1.00	RP	0.65	1.00	RP	0.99	1.00	RP	0.89	0.99
AT1G68862.1	RP	1.00	1.00	RP	1.00	0.81	RP	1.00	0.94	RP	0.56	0.91	RP	0.77	1.00	RP	0.99	1.00	RP	0.89	0.94
AT5G66490.1	RP	1.00	0.84	RP	0.97	0.94	RP	0.99	0.95	RP	0.79	1.00	RP	0.65	0.91	RP	0.93	0.9	RP	0.89	0.92
AT3G42380.1	RP	1.00	0.84	RP	0.99	0.97	RP	0.99	0.89	RP	0.75	1.00	RP	0.84	1.00	RP	0.7	1.00	RP	0.88	0.95
AT2G16881.1	NRP	0.0		RP	1.00	1.00	RP	1.00	0.98	RP	0.89	1.00	RP	0.9	1.00	RP	1.00	1.00	RP	0.8	0.99
AT5G51812.1	NRP	0.0		RP	1.00	1.00	RP	1.00	0.99	RP	0.62	1.00	RP	0.84	1.00	RP	0.97	1.00	RP	0.74	1.00
AT4G08263.2	NRP	0.0		RP	0.93	0.97	RP	1.00	1.00	RP	0.93	1.00	RP	0.94	0.91	RP	0.55	1.00	RP	0.72	0.97
AT2G40711.1	NRP	0.0		RP	0.93	0.9	RP	1.00	1.00	RP	0.59	1.00	RP	0.81	1.00	RP	0.93	1.00	RP	0.71	0.97

Apêndice B

Trabalhos enviados

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.

- 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 Limaz, Anésia A. Santosz, Maximiller D L Costaz, Francisco Murilo Zerbiniz, Fabio R Cerqueira e Elizabeth P B Fontes.
  Evento: CLEI 2016
  Ano: 2016.