UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE CIÊNCIAS
DEPARTAMENTO DE BIOQUÍMICA E BIOLOGIA MOLECULAR
BACHARELADO EM BIOTECNOLOGIA

MATHEUS FINGER RAMOS DE OLIVEIRA

PREDIÇÃO E ANÁLISE DO POTENCIAL CODIFICANTE DE PEQUENAS JANELAS ABERTAS DE LEITURA (sORFs) NO GENOMA DE *JATROPHA CURCAS* L.

FORTALEZA
2022

MATHEUS FINGER RAMOS DE OLIVEIRA

PREDIÇÃO E ANÁLISE DO POTENCIAL CODIFICANTE DE PEQUENAS JANELAS ABERTAS DE LEITURA (sORFs) NO GENOMA DE *JATROPHA CURCAS* L.

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Biotecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Bacharel em Biotecnologia.

Área de concentração: Bioinformática.

Orientador: Prof. Dr. Nicholas Costa Barroso Lima.

FORTALEZA

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

O48p Oliveira, Matheus Finger Ramos de.
Predição e análise do potencial codificante de pequenas janelas abertas de leitura (sORFs) no genoma de Jatropha curcas L. / Matheus Finger Ramos de Oliveira. – 2022.
37 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Ciências, Curso de Biotecnologia, Fortaleza, 2022.
Orientação: Prof. Dr. Nicholas Costa Barroso Lima.

1. Bioinformática. 2. Biocombustível. 3. Biotecnologia. 4. Pinhão-manso. 5. sORFs. I. Título.
CDD 661

MATHEUS FINGER RAMOS DE OLIVEIRA

PREDIÇÃO E ANÁLISE DO POTENCIAL CODIFICANTE DE PEQUENAS JANELAS ABERTAS DE LEITURA (sORFs) NO GENOMA DE *JATROPHA CURCAS* L.

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Biotecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Bacharel em Biotecnologia.

Área de concentração: Bioinformática.

Orientador: Prof. Dr. Nicholas Costa Barroso Lima.

Aprovada em: 01/02/2022.

BANCA EXAMINADORA

_____
Prof. Dr. Nicholas Costa Barroso Lima (Orientador)
Universidade Federal do Ceará (UFC)

_____
Dra. Thais Andrade Germano
Instituto Nacional de Pesquisas da Amazônia (INPA)

_____
Ma. Lyndefania Melo de Sousa
Universidade Federal do Ceará (UFC)

AGRADECIMENTOS

Ao Professor Dr. Nicholas Costa Barroso Lima, por não desacreditar de mim um único segundo e por me ensinar a procurar diferentes caminhos e ideias para resolução de problemas.

Ao Professor Dr. José Helio Costa, pelo acolhimento no seu laboratório e por acreditar no meu potencial e nas minhas habilidades com informática. E aos colegas de pesquisa e amigos Thais, Lyndefânia, Mathias, Edson, Shahid, Susan, Dayane, Talita, e outros, por sempre acreditarem muito em mim, pelos momentos de descontração e por me ajudarem a me encontrar no meio acadêmico.

À Universidade e à Coordenação do Bacharelado em Biotecnologia, pela estrutura e capacitação. Por me permitiram explorar os três pilares de Ensino, Pesquisa e Extensão.

Ao Programa Integrado de Qualificação Discente do Bacharelado em Biotecnologia (PIQD), que despertou meu interesse em divulgar a ciência e me aproximou do meu curso. Aos coordenadores do projeto, que me ensinaram importantes lições e serviram de exemplo de dedicação com a extensão. Aos muitos membros do PIQD, que tornaram minha experiência na universidade inesquecível e foram sempre tão amistosos.

Aos meus familiares, em especial minha mãe (Ivete Noemi), pelo exemplo de garra, força e determinação, e por sempre ter me apoiado em qualquer decisão que eu tomei ao longo da minha vida acadêmica. À minha irmã (Amanda Finger), por me ajudar sempre que pode e por ter me ensinado tudo que sei relacionado a artes. E ao meu pai (Alexandre Ramos), por ter me apoiado e acreditado em mim.

Ao meu grupo de amigos do cluster: Symon, Davi, Ítalo, Agna, João Augusto, Amanda, Bruno, João Neto, Thiago, Júnior, Pathy, Gabriel, Lívia, Romão, Walmick, Ariany, por ter tornado a universidade uma experiência única pra mim e ter enchido meus dias com alegria e leveza.

Aos amigos, colegas de graduação e professores, que gostaria de poder citar todos, por toda a amizade, companheirismo e ensinamentos que permitiram que eu alcançasse lugares e obtivesse conquistas que não seriam possíveis sem o apoio de vocês.

RESUMO

As pequenas janelas abertas de leitura (sORFs) são regiões do genoma entre um códon de início e um códon de parada que são menores ou iguais a 300 pares de base, as quais são removidas da análise dos projetos de genoma devido ao alto número de falsos positivos. Atualmente, diversos trabalhos mostraram a importância dos pequenos peptídeos codificados por sORFs na morfogênese de plantas e animais, porém poucos têm sido identificados e caracterizados. O pinhão-manso (*Jatropha curcas* L.) é uma planta com grande potencial para produção de biodiesel no Brasil, mas ainda há entraves que dificultam a sua adesão pelo mercado, como a ausência de cultivares comerciais e falta de conhecimento sobre as condições nutritivas adequadas para o seu cultivo. Com o objetivo de analisar o potencial codificante de sORFs de *Jatropha curcas* L., foram feitas a predição das sORFs e filtragem por meio da análise de conservação e da busca por domínios proteicos, gerando 446 sORFs não-exônicas com potencial codificante de um conjunto total de 3.734.705 sORFs preditas. Além disso, foi realizada uma análise de enriquecimento de termos de Ontologia Gênica dos domínios presentes nas sORFs, sendo 8 sORFs relacionadas ao processo de biossíntese de lipídios, 1 com relação especificamente ao o processo de biossíntese de ácidos graxos e 8 sORFs com relação ao processo de morfogênese anatômica. Os resultados obtidos poderão ser utilizados para auxiliar em estudos futuros de genômica funcional para compreender melhor a função dessas sORFs de pinhão-manso, provendo informações valiosas para o melhoramento dos caracteres da planta para a produção de biodiesel e outros compostos.

Palavras-chave: bioinformática, biocombustível, biotecnologia, pinhão-manso, sORFs.

ABSTRACT

The small Open Reading Frames (sORFs) are genomic regions between a start codon and a stop codon smaller than 300 base pairs. These regions are generally removed from genome projects annotations due its high false positive rates. However, several studies shown the importance of the small peptides (sPEPs) derived from the sORFs in plant and animal morphogenesis, but the information about their physiological function is still scarce. The physic nut (*Jatropha curcas* L.) is a plant with high potential to the production of biodiesel in Brazil, but there are still a lot of obstacles that hamper its large scale production. There are few comercial cultivars of this species and there is a lack of knowledge about the appropriate nutritional conditions for its growth. To assess the sORFs' coding potential, the prediction and filtering, which consists in the conservation analysis and search for protein domains, were performed, generating 446 novel putative sORFs with coding potential from a total of 3.734.705 initial sORFs. Furthermore, an enrichment analysis of the Gene Ontology terms from the domains from the sORFs was performed, and 8 sORFs were related to the lipid biosynthesis, 1 sORFs were related specifically to the fatty acid biosynthesis, and 8 sORFs were related to the process of anatomic morphogenesis. The results obtained can be used to help future functional genomics research that aim to understand the function of the small ORFs of the physic nut, providing valuable information to the improvement of agronomic features of the cultivars for the production of biodiesel and other products.

Keywords: bioinformatics; biotechonology; biofuel; physic nut; sORFs.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

PB	Processo Biológico
CC	Componente Celular
CDS	*Coding Sequence* – Sequência Codificante
CLE	CLAVATA3/ EMBRYO SURROUNDING REGION
CNPE	Conselho Nacional de Política Energética
FDR	*False Discovery Rate* – Taxa de Falsa Descoberta
FP	Falso Positivo
GO	*Gene Ontology* – Ontologia Gênica
HMM	*Hidden Markov Models* – Modelos Escondidos de Markov
K_A	Substituições Não-Sinônimas
K_S	Substituições Sinônimas
FM	Função Molecular
ML	*Machine Learning*
MME	Ministério de Minas e Energia
NCBI	*National Center for Biotechnology Information* – Centro Nacional para Informação Biotecnológica
ORF	*Open Reading Frame* – Janela de Leitura Aberta
PNPB	Programa Nacional de Produção e Uso de Biodiesel
PV	Positivo Verdadeiro
RIP	*Ribosome-inactivating protein* – Proteína Inativadora de Ribossomo
sORF	*small Open Reading Frame* – Janela de Leitura Aberta pequena
tal	*tarsal-less*

SUMÁRIO

1 INTRODUÇÃO

1.1 *Jatropha curcas* L.: Produção de Biodiesel

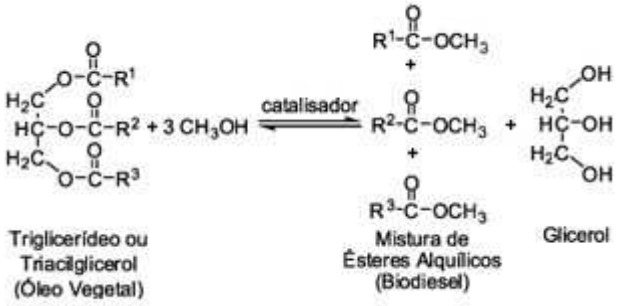
1.1.1 Produção de Biodiesel no Brasil

A demanda por energia tem aumentado nos últimos anos devido ao aumento populacional, à urbanização, e a outras problemáticas sociais. Além disso, fatores como o alto preço do petróleo, mudanças climáticas e poluição do ar têm levado o ser humano a procurar fontes alternativas de energia aos produtos derivados do petróleo (YESILYURT, 2017). Desde a época de 1970, diversos programas de incentivo à produção e uso do biodiesel foram implementados no Brasil. Como exemplo, em 2004 foi criado o Programa Nacional de Produção e Uso de Biodiesel (PNPB), que busca introduzir o biodiesel de forma sustentável na matriz energética brasileira e contribuir para a redução das emissões de gases do efeito estufa (CAVALCANTE FILHO; BUAINAIN; DE SOUZA BENATTI, 2019).

Uma medida feita pelo Ministério de Minas e Energia (MME) e aprovada pelo Conselho Nacional de Política Energética (CNPE) no final de outubro de 2018 prevê um aumento gradual na porcentagem de 10% para 15% v/v de biodiesel no diesel até o final de 2023 (CSOB, 2019).

O biodiesel é um combustível a base de fontes renováveis, como óleo vegetal e gordura animal, sendo quimicamente obtido a partir da transesterificação dos triglicerídeos presentes nessas fontes. Nessa reação orgânica (**Figura 1**) mediada por um catalisador, um éster é transformado em outro através da troca dos grupos alcoóxidos, utilizando um mono-álcool de cadeia curta, como metanol ou etanol (RINALDI *et al*., 2007).

Figura 1 – Esquema da transesterificação de um triglicerídeo genérico com metanol.



Fonte: RINALDI *et al*. (2007).

A cadeia produtiva do biodiesel (**Figura 2**) pode ser dividida em: Produção primária, que consiste nos fornecedores de matéria-prima; Indústrias que transformam a matéria-prima em biodiesel, realizando o esmagamento e a transesterificação; Atacadistas, que são as empresas responsáveis pelo refinamento; Varejistas que distribuem o combustível; E, por fim, o consumidor. Somente 10% das indústrias de esmagamento, que representam a capacidade produtiva, estão presentes nas regiões Norte e Nordeste, que são o foco principal das ações do PNPB (CAVALCANTE FILHO; BUAINAIN; DE SOUZA BENATTI, 2019). Em 2020, a região Nordeste foi responsável por somente 7,43% da produção total de biodiesel no Brasil (ANP, 2021). Uma possível justificativa para esse cenário é o fato de o Nordeste não possuir uma estrutura produtiva consolidada com cadeias agroindustriais instaladas e não ser uma região tradicional do agronegócio brasileiro (CAVALCANTE FILHO; BUAINAIN; DE SOUZA BENATTI, 2019).

Figura 2 – Cadeia produtiva do biodiesel.

Fonte: GOLLO et al. (2010).

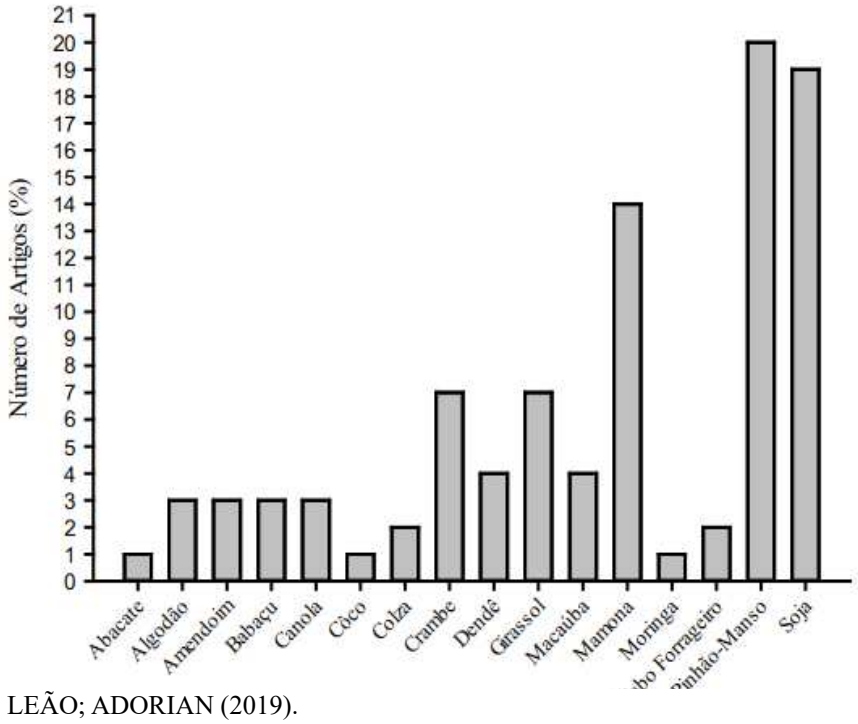
Diversos estudos foram feitos sobre óleos advindos de variadas espécies de plantas oleaginosas objetivando seu uso como biocombustíveis, com a premissa de que essas fontes de energia iriam causar um menor impacto ambiental e que seriam um apoio à agricultura familiar (RAMOS *et al*, 2003).

Em 2020, 71,40% do biodiesel produzido era advindo do óleo de soja (*Glycine max* L.) (ANP, 2021). A grande porcentagem de contribuição da soja como matéria-prima pode ser explicada pela presença de três parâmetros básicos: domínio tecnológico, pois o Brasil é líder nas pesquisas com soja tropical; escala de produção, porque a soja é a única que consegue atender as demandas para produção de biocombustível; e logística, haja visto que a soja é cultivada amplamente em todas as regiões do país (CSOB, 2019).

A grande porcentagem de contribuição da soja para a produção de biodiesel levou a um aumento na devastação de florestas amazônicas, mesmo que o cultivo de soja se concentre em campos do Cerrado. A monocultura de soja, o conflito de terra entre os grandes agricultores de soja com a população mais humilde, entre outros impactos negativos relativos à produção de soja faz com que o balanço ecológico desejado não seja atingido (KOHLHEPP, 2010).

Assim, inúmeras fontes alternativas para a produção do biocombustível já foram analisadas. Em estudo realizado por LEÃO e ADORIAN (2019), realizou-se uma busca pela palavra 'biodiesel' na base de dados *Scientific Electronic Library Online* (SciELO) e o resultado foi filtrado para publicações feitas entre 2005 e 2019. Na **Figura 3**, podemos ver a distribuição das espécies estudadas pelos artigos como matéria-prima para a produção de biocombustível.

Figura 3 – Matérias primas alternativas para produção de biodiesel mais citadas nos artigos analisados, publicados entre os anos de 2005 e 2019, na base de dados SciELO.



Fonte: LEÃO; ADORIAN (2019).

Como uma alternativa à soja, o pinhão-manso (*Jatropha curcas* L.) se mostrou ser uma das espécies mais estudadas, e isso justifica-se pelo seu alto teor de óleo nas sementes e fácil conversão em biodiesel. O incremento do cultivo dessa cultura aumentou a demanda por informações sobre o desenvolvimento da planta, justificando o maior número de estudos com

ela (LEÃO; ADORIAN, 2019).

1.1.2 Potencial do pinhão-manso para a produção de biodiesel

Uma espécie com muito potencial para produção de biodiesel que pode ser facilmente cultivada por pequenos agricultores é a *Jatropha curcas* L., que possui grandes quantidades de óleo em suas sementes. A legalização para a venda das sementes ocorreu em 2008 no Brasil (DE MELO *et al*, 2019).

O pinhão-manso é uma pequena árvore ou grande arbusto diploide de até 5 metros da família *Euphorbiaceae*, possui características de resistência à seca e tem diversos usos medicinais. Suas sementes são tóxicas para humanos, porém eram muito utilizadas para produção de óleos e sabonetes (LAVIOLA *et al*, 2015). Quanto à reprodução da planta, raramente são observadas flores hermafroditas e as flores estaminadas abrem o botão antes das flores pistiladas, promovendo assim a polinização cruzada, a qual é feita geralmente por insetos (HELLER, 1996). O fruto (**Figura 4**) possui diâmetro geralmente entre 1,5 a 3 centímetros e é constituído de 53 a 62% de semente e 38 a 47% de casca (DE ALBUQUERQUE *et al*., 2008).

Figura 4 – Frutos e sementes de *Jatropha curcas* L. em diferentes estágios de maturação.

Fonte: DE ALBUQUERQUE *et al*. (2008).

Essa espécie é exigente em insolação (número de horas nas quais, durante o dia, a radiação solar ocorre sem obstrução de nuvens ou outros fenômenos) e possui forte resistência à seca, crescendo espontaneamente nas regiões mais secas do semi-árido. Por ser uma cultura perene, pode ser utilizada na conservação do solo, pois evita a erosão e a perda de água ao criar uma camada de matéria seca por cima do solo (DE ARRUDA *et al*., 2004).

Como a maioria das sementes de plantas, a semente do pinhão-manso possui uma variedade de compostos tóxicos e antinutricionais, como inibidores de proteases e fitato, também contendo ésteres de forbol e a curcina (HE *et al*, 2011). A ricina, composto similar a curcina, possui maior concentração no endosperma das sementes e tem como função evitar a predação, sendo classificada como Proteína Inativadora de Ribossomo (RIP – *Ribosome-inactivating protein*). Apenas uma molécula de ricina que entra no citosol pode desativar cerca de 1500 ribossomos por minuto (DA SILVA, SOTO-BLANCO, 2014).

Foram relatados genótipos chamados de "não-tóxicos" que não apresentam ésteres de forbol. No entanto, esses genótipos ainda possuem diversos compostos antinutricionais e não foi estabelecido se os níveis de curcina são menores que nos demais genótipos (HE *et al*, 2011). O aumento da produção de *J. curcas* L. no Brasil pode acarretar em uma maior taxa de acidentes que levam à intoxicação com curcina. Logo, é interessante tomar medidas para reduzir os riscos de intoxicação, como utilização de silenciamento gênico para produção de cultivares que não produzam a toxina, sendo necessário se obter conhecimento dos genes envolvidos na sua biossíntese.

O percentual de germinação do pinhão-manso varia entre 60 a 80% e, em produção econômica, a espécie vai até 20 a 25 anos, quando a produtividade decai drasticamente e é recomendável substituir por outra cultura. A produção de óleo é pequena inicialmente, porém aumenta ao longo das sucessivas safras, até a estabilização entre os 5 e 6 anos iniciais (DE MELO *et al*., 2019). É imprescindível aumentar a taxa de germinação e otimizar o tempo de produção do pinhão-manso para torná-lo competitivo no mercado.

Como principais vantagens, a *J. curcas* L. não é utilizada na alimentação, diferente da soja, reduzindo mudanças de preço de acordo com a demanda alimentar. Entretanto, a espécie ainda precisa de cultivares com caracteres agronômicos desejáveis e não se conhece ainda as melhores condições para o cultivo da espécie. Pouco se sabe ainda sobre sua base genética, não havendo informações também sobre o genoma das cultivares utilizadas no Brasil (DE MELO *et al*, 2019). Até 2015, não haviam estudos com cultivares comerciais do pinhão-manso, o que compromete a uniformidade dos resultados anteriores obtidos por estudos com cultivares selvagens, devido sua grande variação genética (PEREIRA *et al*, 2018). Nesse contexto, o desenvolvimento da planta deve ser estudado mais detalhadamente, principalmente no âmbito molecular.

Atualmente, os principais empecilhos para o pinhão-manso não ter se tornado amplamente difundido como matéria-prima para o biodiesel são, além da falta de padronização devido a falta de cultivares comerciais, a falta de conhecimento sobre os

requerimentos nutricionais da planta, alta variação na quantidade e qualidade das sementes, alta taxa de flores macho em comparação como flores fêmea, toxicidade da semente e estresses abióticos (MAZUMDAR *et al*, 2018).

1.2 Histórico de pesquisa com sORFs

Open Reading Frame (ORF), do inglês - Janela de Leitura Aberta, é um conceito muito utilizado nos campos da Genômica e da Bioinformática para identificação de *Coding Sequences* (CDS), do inglês - Sequências Codificantes, sendo frequentemente descrita como uma região entre um códon de início da tradução (ATG) e um códon de parada (TGA/TAG/TAA) cujo tamanho é divisível por 3. Entretanto, essa descrição foi criada quando a maioria dos genomas sequenciados eram de seres procariotos, que majoritariamente não possuem íntrons, e toda ORF era uma potencial CDS. Em eucariotos, essa definição não é tão adequada devido à presença de íntrons e *splicing*, sendo utilizadas definições que delimitam as ORFs por códons de parada, ou ainda éxons com potencial codificante identificadas por algoritmos de anotação gênica (SIEBER; PLATZER; SCHUSTER, 2018).

Os algoritmos de anotação funcional de genomas têm como objetivo distinguir ORFs que traduzam proteínas de sequências aleatórias, e para isso muitas vezes aplicam pontos de corte de tamanho mínimo. A probabilidade uma sequência ser aleatória é maior em sequências pequenas, o que faz com que potenciais pequenas ORFs codificantes sejam menosprezadas pelos algoritmos atuais, sendo classificadas com sequências não-codificantes. *small* ORFs (sORFs) são ORFs cujo tamanho é menor do que o mínimo para detecção para tais algoritmos, o qual geralmente é em torno de 300 pares de bases (MAKAREWICH; OLSON, 2017).

Estudos recentes que utilizam de novas tecnologias, como o *Ribosome profiling* (Ribo-Seq) e análises proteômicas adaptadas, mostraram que sequências antes categorizadas como não-codificantes eram traduzidas (BRUNET; ROUCOU, 2019; DELCOURT *et al*, 2018; ERHARD *et al*, 2018; MA *et al*, 2016; OLEXIOUK; VAN CRIEKINGE; MENSCHAERT, 2018). Parte dessas proteínas eram micropeptídeos originados de sORFs ignoradas pelos algoritmos de anotação, o que levou a uma série de publicações voltadas para estudos dessas pequenas sequências codificantes (BRUNET; LEBLANC; ROUCOU, 2020).

Uma das primeiras sORFs a serem caracterizadas foi o gene *tarsal-less* (*tal*), descoberto ao se estudar a mutação espontânea em que indivíduos do gênero *Drosophilla* não desenvolviam os segmentos do tarso e, logo, possuíam pernas defeituosas (GALINDO *et al*,

2007). Ao fazer uma análise de genes que eram up-regulados durante o crescimento, descobriram o transcrito do gene *tal*, anteriormente anotado como não-codificantes, mas que codificava três peptídeos de 11 a 32 aminoácidos (GALINDO *et al*, 2007).

Pode-se considerar que o primeiro estudo sistemático de sORFs foi uma pesquisa feita por Kanstermayer e colaboradores, que realizaram uma busca em estudos anteriores e encontraram um total de 299 sORFs não-anotadas em *Saccharomyces cerevisiae*. Além disso, também foram feitos experimentos com deleções desses genes e, assim, foi possível relacioná-los com funções importantes no metabolismo da levedura (KANSTERMAYER *et al*., 2006). Em plantas, um dos primeiros estudos envolvendo sORFs foi realizado em *Arabidopsis thaliana*, em que foram encontradas 3.241 sORFs intergênicas com potencial codificante, posteriormente sendo observado que 49 destas induziam um forte fenótipo quando superexpressas. (HANADA *et al*, 2007, 2013).

Atualmente, vários estudos comprovam que produtos de ORFs menores que 300 nucleotídeos possuem papel fundamental em várias vias metabólicas presentes nos organismos, sendo expressos durante a morfogênese de plantas e animais. Em plantas, as sORFs podem, por exemplo, regular a formação da raiz, controlar o formato da folha, e atuar na divisão de células do câmbio cortical durante a formação de nódulos. Porém, ainda há poucos estudos abordando o papel desses pequenos peptídeos, tendo em vista que é uma área que recebeu mais destaque somente nos últimos anos (ALBUQUERQUE *et al*, 2015).

Um exemplo de peptídeos codificados por sORFs importantes para o desenvolvimento de plantas são os da classe CLAVATA3/ EMBRYO SURROUNDING REGION (CLE). As enzimas da CLE atuam na manutenção de meristemas de raiz, caule e flor, na emergência de raízes laterais e no desenvolvimento vascular (DE CONINCK; DE SMET, 2016).

1.3. Métodos para descoberta de novas sORFs

Janelas de leitura aberta são definidas como a região delimitada por um códon de início de um códon de término e, em projetos de genoma de procariotos, potencialmente constituem regiões codificantes de proteínas. A predição de ORFs *ab initio* é muito comum em genomas de procariotos e utiliza algoritmos que analisam sinais, como regiões TATA box e regiões importantes para o recrutamento de ribossomos, e conteúdo, como regiões conservadas entre espécies, e que fazem uso de redes neurais, transformadas de Fourier e modelos de Markov (VERLI, 2014).

A análise do potencial codificante de sORFs pode ser dividida em três principais abordagens: predição computacional, experimentos com Ribo-Seq e utilização de espectrometria de massa (CHUGUNOVA *et al*, 2018). A técnica Ribo-Seq consiste na adição de nucleases de RNA para degradar as sequências de nucleotídeos que não estão protegidas pela ligação ao ribossomo, seguida pelo isolamento e sequenciamento dos mRNAs associados aos ribossomos (CALVIELLO, OHLER, 2017). A espectrometria de massa funciona aplicando cromatografia líquida seguida por espectrometria de massa em tandem, onde se obtém a massa de peptídeos e fragmentos de proteínas hidrolisados, a qual é posteriormente checada contra bancos de dados para a identificação das proteínas (CHUGUNOVA *et al*, 2018).

A descoberta de sORFs usando métodos computacionais pode ser feita por meio de buscas usando o BLAST contra genomas de outras espécies, técnica que depende da homologia com sequências já depositadas nos bancos de dados, ou utilizando métodos *ab initio*, que consideram características típicas de sequências codificantes, como preferência de códons de sequências codificantes, transformação das sequências em matrizes, Modelos Escondidos de Markov (*Hidden Markov Models* - HMM) ou modelos híbridos de HMM com máquina de vetores de suporte, e também viés de composição hexamérica (CHENG *et al*, 2011).

Atualmente, é possível encontrar diversas abordagens em algoritmos para diferenciar sequências codificantes de não-codificantes, porém é possível notar algumas características que são geralmente utilizadas para tal: a identificação de sequências conservadas entre as espécies; pesquisa por viés de uso de códons; e avaliação da similaridade da sequência a outras proteínas ou domínios depositados em bancos de dados (CHUGUNOVA *et al*, 2018).

Técnicas baseadas em *Machine Learning* (ML) têm sido cada vez mais utilizadas nos últimos anos e consistem em um conjunto de algoritmos utilizados para criar um modelo que reconheça padrões a partir de um conjunto de dados e gerar um resultado de classificação, *clusters*, entre outros. Diversos softwares de bioinformática utilizam ML e, com a grande quantidade de dados experimentais de Ribo-Seq e Espectrometria de Massa disponíveis atualmente, ferramentas que utilizam ML a partir desses dados têm uma boa taxa de predição correta de sORFs codificantes. No presente, temos como principais softwares de análise de sORFs codificantes o sORFfinder e o MiPepid (ZHU, GRIBSKOV, 2019).

Para facilitar a compreensão da função de um grande número de sequências, muitas vezes se utiliza o banco de dados de Ontologia Gênica (*Gene Ontology* – GO), o qual

detém termos genéricos para classificar a funcionalidade de sequências, genes ou produtos gênicos. Os termos são estruturados em uma classificação que possui relações de ‘é um’ e ‘pertence à’. Existem três principais termos não-sobrepostos em que todas as sequências são classificadas: Processos Biológicos, Função Molecular e Componentes Celulares (GENE ONTOLOGY CONSORTIUM, 2004).

Devido ao grande número de termos GO, é necessário utilizar categorias mais gerais. Cada termo GO possui um nível hierárquico e os termos mais genéricos aos quais cada um pertence podem ser rastreados, o que levou a criação dos subconjuntos de termos GO, ou GO *slims*. Esses subconjuntos juntam termos relevantes para um problema ou conjunto de dados específico, como ‘lipídios’ ou ‘fotossíntese’. Os GO *slims* podem ser gerais, específicos para domínios (como o subconjunto de GO para plantas) ou específicos proteínas (DAVIS; SEHGAL; RAGAN, 2010).

2 OBJETIVOS

2.1. Objetivo geral

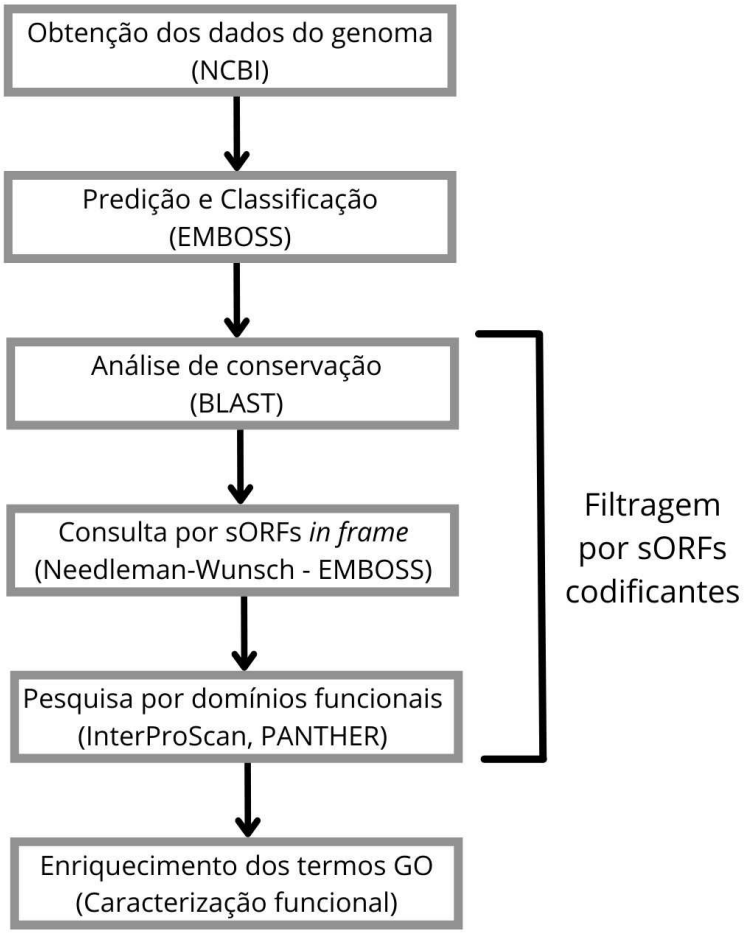
- Predição de Janelas Abertas de Leitura curtas (sORFs) em pinhão-manso (*Jatropha curcas* L.), bem como a análise do potencial codificante e caracterização funcional destas.

2.2. Objetivos Específicos

- Predição de sORFs de *Jatropha curcas* L;
- Análise da conservação das sORFs preditas contra espécies filogeneticamente próximas;
- Análise das sORFs *bona fide* com MiPepid;
- Análise da ontologia gênica dos domínios presentes nas sORFs com potencial codificante.

3 METODOLOGIA

Figura 5 – Representação da Metodologia do presente trabalho em forma de fluxograma.



Fonte: Imagem elaborada pelo autor.

3.1 Predição e classificação das sORFs

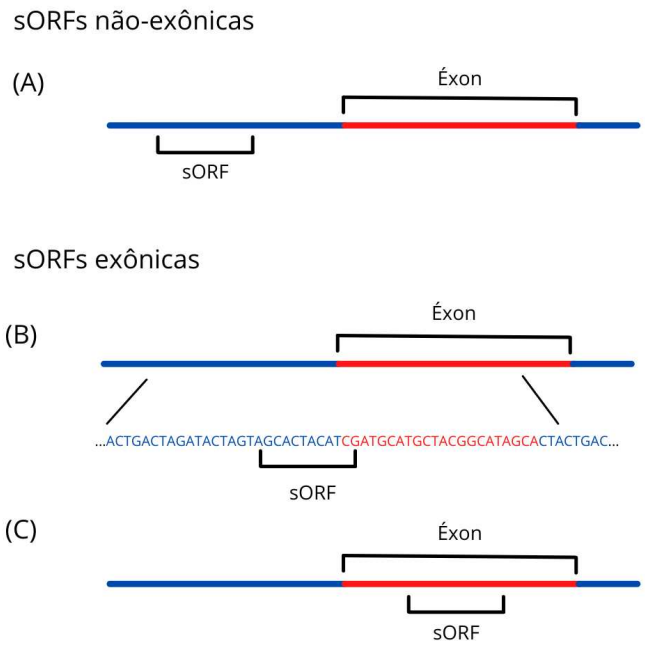
O genoma do pinhão-manso foi recuperado no formato fasta do projeto de montagem com número de acesso GCA_000696525.1, que trabalhou com a cultivar GZQX0401 de *J. curcas* L. (ZHANG *et al*, 2014). Além do arquivo do genoma no formato fasta, também foi utilizado o arquivo de anotação no formato GFF do mesmo projeto de montagem. Ambos os arquivos foram baixados por meio do NCBI (*National Center for Biotechnology Information*) (https://www.ncbi.nlm.nih.gov/).

Para obter-se todas as possíveis sORFs dentro do genoma de *J. curcas*, utilizou-se a função *getorf* do EMBOSS versão 6.6.0.0 (RICE, BLEASBY, 2000), considerando as ORFs que começavam em ATG e terminavam em TAA, TGA ou TAG, possuíam no máximo 300

nucleotídeos e no mínimo 30, sem contar com o códon de parada.

Com base na anotação no formato GFF, foi possível classificar as sORFs em não-exônicas e exônicas considerando a região gênica onde se encontram. As sORFs foram classificadas como não-exônicas quando nenhum nucleotídeo de sua sequência estava inserido em algum éxon anotado (**Figura 6a**), enquanto as exônicas são as sORFs que estão parcialmente (**Figura 6b**) ou totalmente (**Figura 6c**) inseridas dentro de éxons anotados. As sORFs não-exônicas são o foco do trabalho pois representam as sORFs inéditas, enquanto as exônicas representam sORFs que já foram anotadas ou que são somente fragmentos de um éxon. Porém, as sORFs exônicas também são importantes para comparação porque representam sequências codificantes fidedignas.

Figura 6 – Classificação das sORFs de acordo com os éxons presentes no arquivo de anotação.



(A) Representação de uma sORFs não-exônica, sem nenhum nucleotídeo presente em uma região anotada como éxon. (B) Exemplo de sORF exônica com parte da sequência presente na região anotada como éxon. (C) Exemplo de sORFs exônica com toda a sequência dentro de uma região anotada como éxon. Fonte: Imagem elaborada pelo autor.

3.2 Cálculo do FDR e análise de conservação

Foi realizada uma análise de conservação dessas sORFs entre espécies da mesma família (*Euphorbiaceae*), utilizando a função tblastn do software blast+ versão 2.5.0. Os

genomas utilizados na análise foram das espécies *Ricinus communis* L., *Hevea brasiliensis* L. e *Manihot esculenta* Crantz, com os respectivos números de acesso do GenBank dos projetos de montagem usados: GCA_000151685.2, GCA_010458925.1 e GCA_013618965.1.

Para decidir se a sORF foi conservada ou não, foi utilizado o valor de *e-value* do alinhamento, e o ponto de corte foi estabelecido por meio de uma análise de Taxa de Falsas Descobertas (*False Discovery Rate* - FDR). Para essa análise, foram criados como controle negativo 54.000 sORFs não-exônicas reversas, ou seja, que começam com códon de parada e terminam com códon de início, a partir do genoma, e 54.000 fragmentos de éxons como controles positivos, a partir do arquivo de anotação. Os tamanhos dos fragmentos de éxons (controle positivo) foram selecionados a partir do tamanhos das sORFs preditas para se obter uma distribuição similar. Essa análise é importante pois analisa a probabilidade de ocorrerem eventos aleatórios que possam erroneamente parecer significantes, como o alinhamento aleatório de sequências não-conservadas.

O controle negativo e o controle positivo foram então traduzidos para proteína, por meio da função *translate* da biblioteca Biopython versão 1.78, e alinhados contra espécies da mesma família (*Euphorbiaceae*), utilizando a função *tblastn* do software blast+ versão 2.5.0. Os genomas utilizados na análise foram das espécies *Ricinus communis* L., *Hevea brasiliensis* L. e *Manihot esculenta* Crantz, com os respectivos números de acesso do GenBank dos projetos de montagem usados: GCA_000151685.2, GCA_010458925.1 e GCA_013618965.1. Então o valor do FDR foi calculado para o *e-value* selecionado com a fórmula FP/(PV + FP), sendo FP os falsos positivos e PV os positivos verdadeiros (LADOUKAKIS *et al*, 2011).

Após a escolha do ponto de corte usando o resultado da análise de FDR, as sORFs preditas foram alinhadas contra os mesmos genomas utilizando a função *tblastn* para avaliar quais as sORFs conservadas.

3.3 Seleção de sORFs *in frame*

Devido ao fato da função *tblastn* realizar um alinhamento local, o qual alinha somente partes da sequência que são similares, foram recuperados a sequência na qual as sORFs consideradas como conservadas alinharam, 300 nucleotídeos antes do começo da sequência e 300 nucleotídeos após do final da sequência, assim cobrindo completamente as regiões 3’ e 5’ da sORF que talvez não tenha alinhado. Então, a sequência resgatada foi realinhada às sORFs utilizando a função *needle* do EMBOSS versão 6.6.0.0, que realiza o

alinhamento global Needleman-Wunsch em pares. No resultado do alinhamento, foi verificado se os códons de início e de parada estavam alinhados e se estavam in frame, ou seja, dentro de uma mesma janela de leitura.

3.4 Avaliação de sORFs com domínio funcional e enriquecimento dos termos de Ontologia Gênica relacionados

Para analisar quais produtos proteicos das sORFs possuíam domínios funcionais, utilizou-se o software InterProScan versão 5.52-86.0 para realizar a busca por similaridade contra o banco de dados curado PANTHER, o qual possui famílias e subfamílias de genes e proteínas relacionadas de acordo com a sua funcionalidade.

O programa InterProScan também foi utilizado para recuperar o id GO relacionado a cada domínio encontrado. Foi utilizado a pacote *GO.db* (CARLSON *et al*, 2017) versão 3.10.0 na plataforma RStudio versão 1.2.5042 para recuperar os termos GO associados a cada id GO recuperado. Também foram analisados os GO *slims* de plantas associadas a cada termo de cada categoria (BERARDINI *et al*, 2004), por meio da tabela no formato OBO (disponível no site http://geneontology.org).

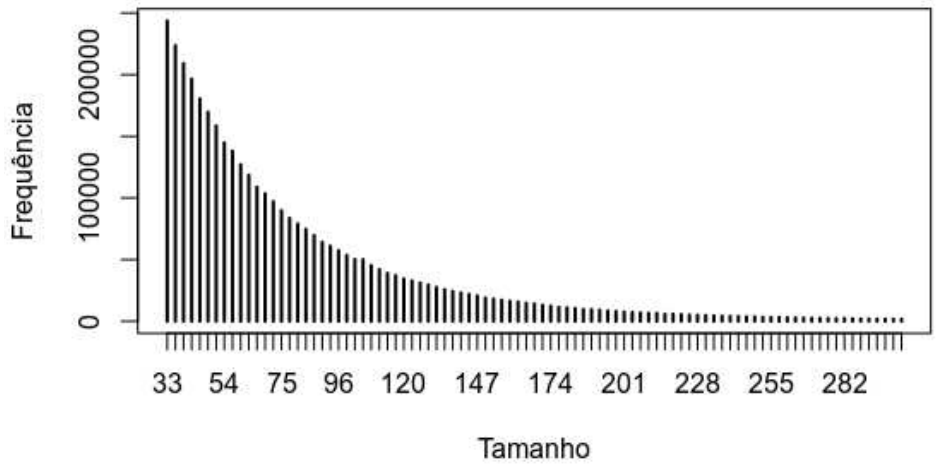
3.5 Análise das sORFs *bona fide* com MiPepid

Em razão de ser um software atual e com boa taxa de previsão correta de sORFs codificantes, além de necessitar de pouco poder de processamento, o software MiPepid foi utilizado para identificar as sORFs com potencial codificante. O algoritmo desse programa utiliza a técnica de *Machine Learning*, em que um modelo de regressão logística é utilizado para calcular a probabilidade das sORFs serem codificante de acordo com os padrões de nucleotídeos de sequência. (ZHU, GRIBSKOV, 2019). O programa foi utilizado em todas as etapas de seleção de sORFs para avaliar o potencial codificante ao longo da filtragem.

4 RESULTADOS E DISCUSSÃO

As sORFs por muito tempo foram ignoradas devido à sua alta probabilidade de serem geradas ao acaso e se acreditava que essas sequências não possuíam funções importantes no metabolismo. Entretanto, já foi evidenciada a importância dessas pequenas janelas de leitura em várias espécies de seres vivos, e várias estratégias estão sendo criadas ou adaptadas para a detecção de sORFs (ORR *et al*, 2020). No atual estudo, inicialmente foram obtidas 3.734.705 sORFs putativas cuja frequência dos tamanhos pode ser observada no **Gráfico 1**. É possível observar que dentre as sORFs iniciais, a maioria possui tamanho próximo a 30 nucleotídeos, que foi o ponto de corte selecionado para o tamanho mínimo das sORFs desse estudo. Isso acontece pois quanto maior a sequência, maior a probabilidade de um haver um códon de parada ao acaso, o que não acontece em sequências menores. A maioria dessas sequências pequenas não possui significado biológico e aparece aleatoriamente (BASRAI *et al*, 1997).

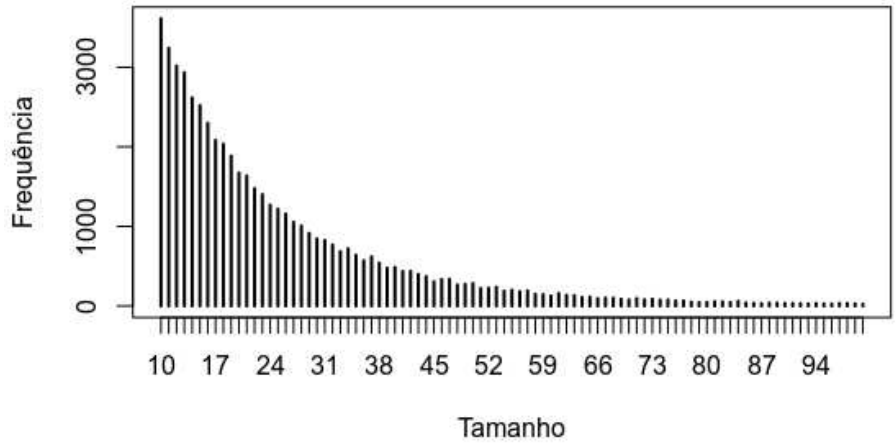
Gráfico 1 – Tamanho das sORFs iniciais em nucleotídeos versus sua frequência.



Fonte: Elaborado pelo autor.

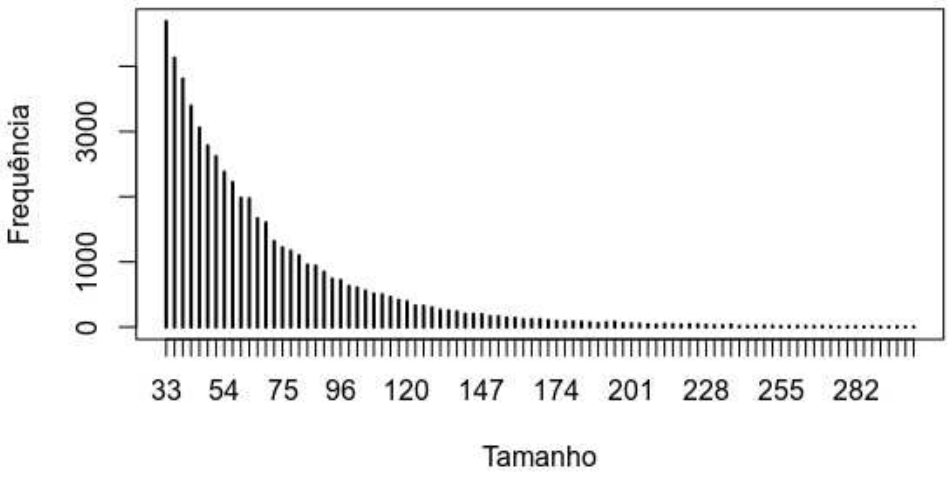
Ambos os controles positivo (fragmentos de éxons) e negativo (sORFs reversas) apresentaram a mesma distribuição de tamanho das sORFs preditas (**Gráficos 2 e 3**), estando aptos para a análise de FDR. Uma distribuição de tamanhos diferentes poderia influenciar no resultado do Blast, pois regiões maiores têm mais chances de alinhar com sequências aleatórias.

Gráfico 2 – Tamanho dos fragmentos de éxons (controle positivo) em aminoácidos versus sua frequência.



Fonte: Elaborado pelo autor.

Gráfico 3 – Tamanho das sORFs reversas (controle negativo) em nucleotídeos versus sua frequência.



Fonte: Elaborado pelo autor.

O ponto de corte escolhido para o *e-value* foi 0,05, que obteve como maior valor 0,132% na análise de FDR para cada genoma alinhado (**Tabela 1**). Assim, 465.675 sORFs (12,47%) foram consideradas como conservadas porque alinharam com, pelo menos, um dos genomas das espécies da mesma família com o valor de *e-value* inferior ou igual ao citado.

Tabela 1 – Valor do FDR obtido para cada espécie calculado a partir do número de positivos e falsos positivos.

Espécie	Proporção positivos verdadeiros	Proporção falsos positivos	FDR (%)
Ricinus communis L.	0,519000	0,000685	0,131846
Manihot esculenta Crantz	0,505759	0,000481	0,095109
Hevea brasiliensis L.	0,500463	0,000574	0,114577

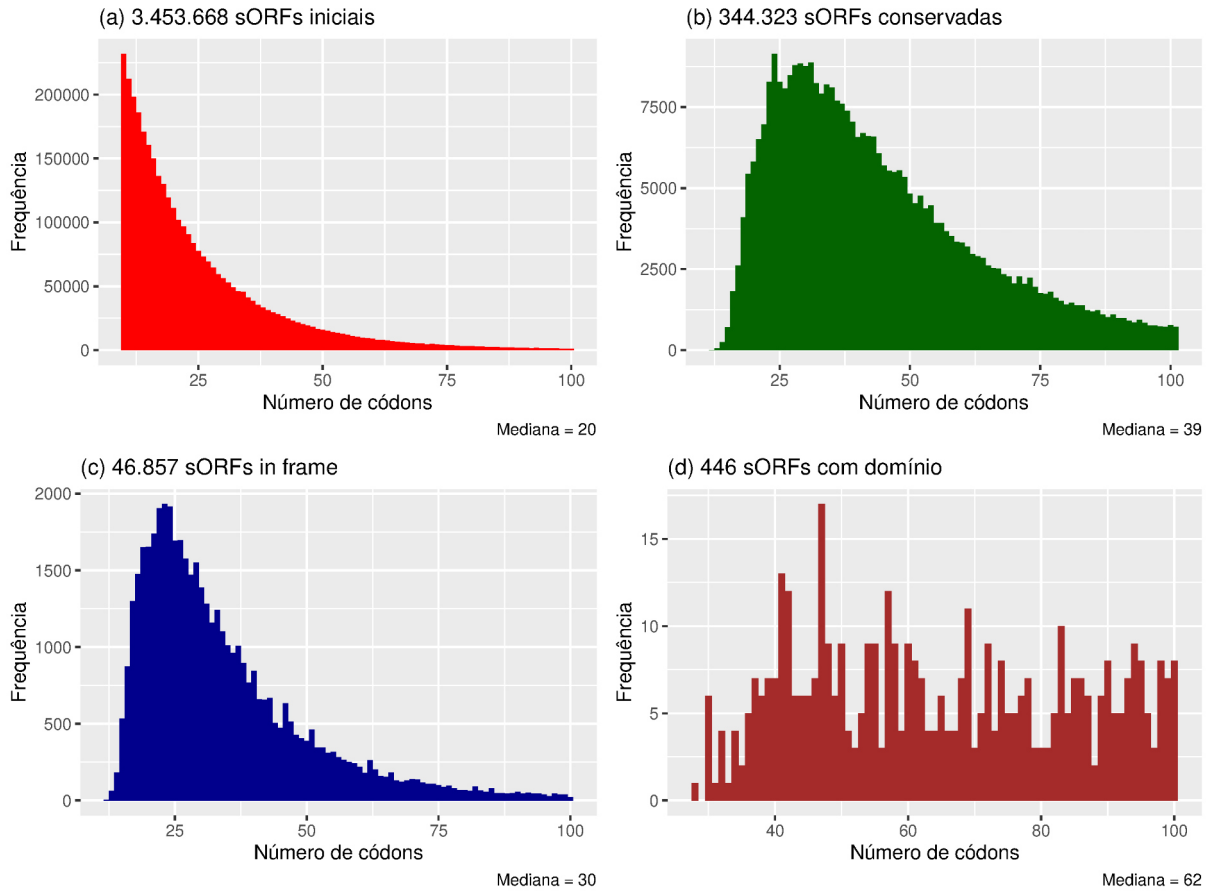
Fonte: elaborada pelo autor.

A seleção das sORFs *in frame* resultou em 73.627 sORFs que compartilhavam códons de início e de parada, ao invés de compartilhar somente um desses códons ou somente uma parte pequena da sequência, sem formar uma ORF.

No resultado da análise dos domínios funcionais, 3.395 sORFs apresentaram pelo menos um domínio funcional. Para abreviar os filtros aplicados nessas sORFs, será utilizado o termo ‘sORFs confiáveis’ para se referir a elas.

A tendência de sORFs menores terem maior chance de acontecer ao acaso é também reforçada pela mediana do tamanho das sORFs, que aumenta após cada etapa de filtragem (**Gráfico 4**).

Gráfico 4 – Distribuição da frequência do tamanho das sORFs não-exônicas em cada etapa do processo.



Fonte: Elaborado pelo autor. Gráfico criado por meio dos pacotes *ggplot2* (WICKHAM *et al*, 2009) e *cowplot* (WILKE, 2020) no software RStudio.

Quanto à classificação, a maioria (92,47%) das sORFs iniciais foram consideradas não-exônicas. Isso deve-se ao fato de que a região ocupada por éxons é pequena em comparação ao resto do genoma, logo a chance de um códon de início e um códon de parada acontecerem ao acaso é menor em éxons. Entretanto, quando analisamos o potencial codificantes das sORFs preditas, a maioria das consideradas como codificantes está na região dos éxons: somente 13,14% das sORFs consideradas confiáveis foram classificadas como não-exônicas, enquanto as outras sORFs possuíam pelo menos parte de sua sequência dentro de um éxon anotado. A maioria das sORFs consideradas codificantes são, na realidade, fragmentos de sequências codificantes por estarem na região exônica. Assim, parte de sua sequência irá ser traduzida. Uma porcentagem maior de sORFs exônicas, ou seja, que terão pelo menos parte de sua sequência traduzida, consideradas como codificantes demonstra que os filtros utilizados para a seleção de sequências codificantes foram efetivos.

Do total de sORFs conservadas, 400.092 (85,92%) foram consideradas como codificantes pelo software MiPepid. Enquanto que, considerando todas as sORFs, o programa classifica como codificante 2.518.338 sORFs (67.43%). Quando o software foi utilizado para analisar as sequências que passaram pelo filtro que analisava os códons de início e de parada dentro da janela de leitura, 65.975 (89,61%) de um total de 73.627 sORFs foram consideradas como codificantes pelo programa e, em relação às sORFs confiáveis, todas as 3.395 sORFs foram consideradas como codificantes pelo MiPepid (**Tabela 2**), ressaltando que as sORFs conservadas e com a presença de domínios tem maior chance de serem transcritas e traduzidas.

Tabela 2 – Número de sORFs após cada etapa de filtragem, assim como a porcentagem de sORFs não-exônicas e consideradas como codificantes pelo software MiPepid a cada etapa.

Etapa	sORFs totais	sORFs não-exônicas	sORFs totais consideradas como codificantes pelo software MiPepid
Predição das sORFs	3.734.705	3.453.668 (92,47%)	2.518.338 (67,43%)
Análise da conservação	465.675	344.323 (73,94%)	400.092 (85,92%)
Avaliação de sORFs *in frame*	73.627	46.857 (63,64%)	65.975 (89,60%)
sORFs com domínio funcional	3.395	446 (13,14%)	3.395 (100%)

Fonte: elaborada pelo autor.

Um estudo feito com *Arabidopsis thaliana* (HANADA *et al*, 2007) que buscava por sORFs na região intergênica utilizando indícios de transcrição de conservação chegou a conclusão de que cerca de 941 das sORFs preditas são de fato codificantes, o que representa cerca de 5% do total de genes. Cerca de 5% das sORFs preditas em *Mus musculus* em outro estudo também foram consideradas como codificantes (FRITH *et al*, 2006). Hanada e colaboradores sugeriram que esse padrão comum de 5% dos genes pode ser compartilhado por espécies diferentes, mesmo que distantes filogeneticamente.

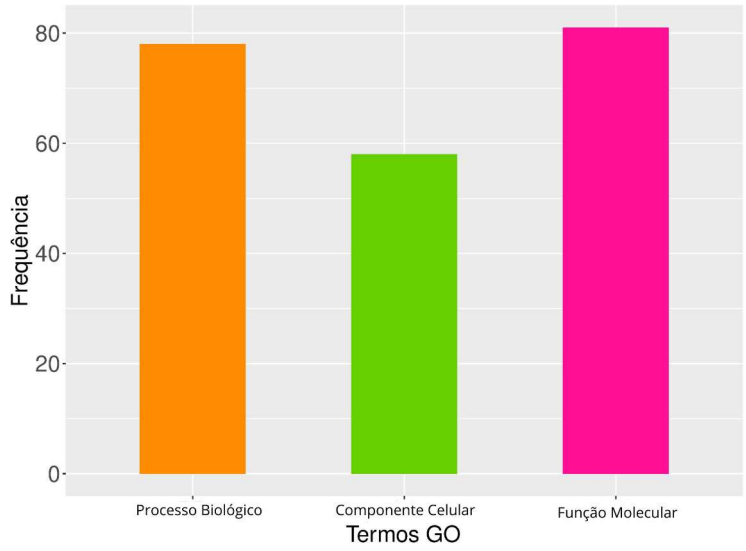
Em comparação com o obtido pelas análises feitas, em que as 446 sORFs não-exônicas representam cerca de 1,09% do total de genes anotados de *J. curcas* L., há uma pequena disparidade, provavelmente devido aos diferentes métodos utilizados. Diferente das análises realizadas em *A. thaliana* e *M. musculus*, não foi realizada uma análise da taxa K_A/K_S para a conservação, mas sim um alinhamento com tBLASTn e análise dos códons de início e

parada *in frame*. Também não foram analisados indícios de transcrição, mas uma busca por domínios funcionais.

Em um estudo feito com *Drosophila pseudoobscura* (LADOUKAKIS *et al*, 2011), os autores chegaram a uma estimativa de 395 sORFs codificantes não-exônicas após análise similar à que foi realizada no presente estudo, valor próximo às 446 sORFs não-exônicas consideradas como codificantes em *J. curcas* L.

A partir do conjunto de 446 sORFs não-exônicas confiáveis, foi possível recuperar termos GO associados aos domínios encontrados de 96 sORFs. Na maioria dos casos, múltiplos termos foram associados à mesma sORF. 78 (35,94%), 81 (37,33%) e 58 (26,73%) termos foram enriquecidos com ORFs relacionadas a Processos Biológicos (PB), Função Molecular (FM) e Componentes Celulares (CC) respectivamente (**Gráfico 5**).

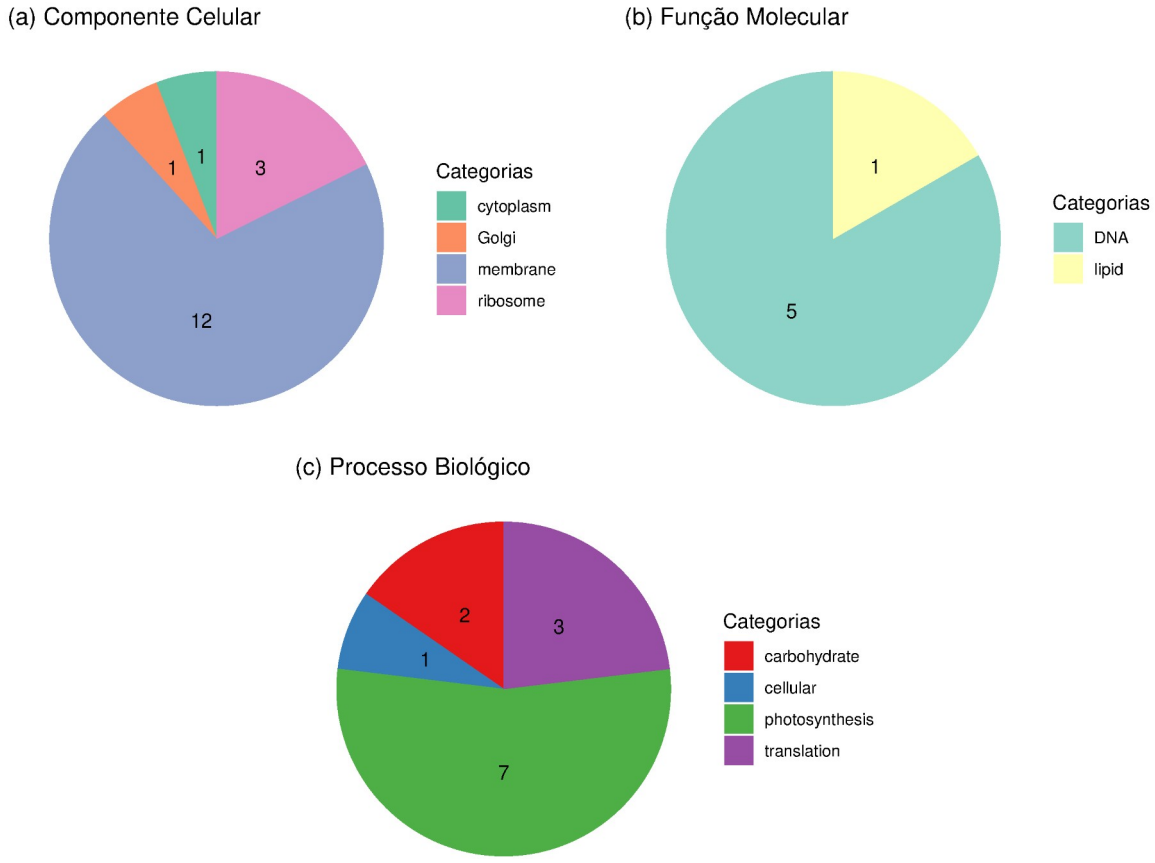
Gráfico 5 – Gráfico de barras dos termos GO presentes nas sORFs confiáveis não-exônicas.



Fonte: Elaborado pelo autor. Gráfico criado por meio do pacote *ggplot2* (WICKHAM *et al*, 2009) no software RStudio.

Então, os ids GO foram utilizados para recuperar os 36 subconjuntos GO associados, dos quais 13, 17 e 6 subconjuntos estavam relacionadas a PB, CC e FM respectivamente (**Gráfico 6**). Quanto às categorias associadas à PB, 7 são ligadas à Fotossíntese, 3 à Tradução, 2 a Carboidratos e 1 a Processos Celulares. Em relação à FM, existiam 5 ligadas ao DNA e 1 ligada a lipídios. Sobre as categorias ligadas a CC, 12 tinham ligação com a Membrana, 3 com o ribossomo, 1 com o Citoplasma e 1 com o Complexo de Golgi.

Gráfico 6 – Gráficos de pizza dos subconjuntos GO presentes nas sORFs confiáveis não-exônicas.



Fonte: Elaborado pelo autor. Gráfico criado por meio dos pacotes *ggplot2* (WICKHAM *et al*, 2009) e *cowplot* (WILKE, 2020) no software RStudio.

Dentre as sORFs confiáveis não-exônicas, 8 possuíam domínios de famílias do PANTHER associadas ao processo de biossíntese de lipídios (GO:0008610), 1 com relação o processo metabólico de ácidos graxos especificamente (GO:0006631) e 8 com relação à morfogênese de estrutura anatômica (GO:0009653).

Em *A. thaliana*, os peptídeos codificados por micro-RNAs miPEP164a, miPEP165a e miPEP319a que atuam na morfogênese da planta mostraram resultados positivos em termos de tamanho do caule, floração antecipada, aumento no número de flores e tamanho do talo floral em comparação ao controle quando aplicados de forma exógena por meio de spray, irrigação, compostagem e adição de fertilizante (COMBIER; LAURES-SERGUES; BECARD, 2020). Isso demonstra sORFs envolvidas com a morfogênese podem influenciar também na floração, que é um problema no pinhão-manso devido a alta taxa de

flores macho em comparação com flores fêmea.

Ademais, os resultados aqui apontados precisam de confirmação a partir de estudos de proteômica, como espectrometria de massa e RiboSEQ, e outros experimentos. Após a validação, esses dados podem ser utilizados para o melhoramento genético de cultivares de pinhão-manso, por meio da criação de marcadores para seleção, CRISPR ou outras técnicas, possibilitando que a espécie seja uma importante fonte de óleo para a produção de biocombustíveis ou outros produtos de alto valor agregado.

5 CONSIDERAÇÕES FINAIS

Os resultados obtidos indicam a presença de várias sORFs em *Jatropha curcas* L. com potencial codificante, as quais podem possuir papel-chave em vias metabólicas alvo de melhoramento genético para tornar a produção de biodiesel e outros compostos a partir do pinhão-manso viáveis economicamente, como a via metabólica de produção de ácidos graxos e a via de produção de toxinas. Adquirir caracteres agronômicos desejáveis irá permitir que o pinhão-manso se torne competitivo, fornecendo outras opções para a produção de biodiesel além da soja.

Além da importância para a produção do biodiesel, os resultados aqui obtidos sobre sORFs de pinhão-manso, após validação experimental, fomentam bancos de dados com anotação de sORFs e seus produtos proteicos. Esses dados podem ser utilizados para auxiliar a treinar programas que utilizam *Machine Learning*, tornando-os mais robustos, assim auxiliando na descoberta de pequenos peptídeos em outras espécies.

Os próximos passos são a anotação das sORFs a partir de dados de proteômica e de Ribo-Seq, análise dos genes e peptídeos individuais e, então, estudos experimentais para testar a função desses genes na fisiologia da planta. Os resultados obtidos serão importantes para a análise experimental da função das sORFs, haja visto que um estudo experimental geralmente analisa poucas sequências, e a seleção de ORFs com potencial codificante facilita a escolha da sequência a ser estudada.

REFERÊNCIAS

ALBUQUERQUE, João Paulo *et al*. small ORFs: a new class of essential genes for development. **Genetics and molecular biology**, v. 38, n. 3, p. 278-283, 2015.

ANP, Agência Nacional do Petróleo, Gás Natural e Biocombustíveis. **Anuário Estatístico 2021**. Disponível em: <https://www.gov.br/anp/pt-br/centrais-de-conteudo/publicacoes/anuario-estatistico/anuario-estatistico-2021>. Acesso em: 13 fev. 2022.

BASRAI, Munira A.; HIETER, Philip; BOEKE, Jef D. Small open reading frames: beautiful needles in the haystack. **Genome research**, v. 7, n. 8, p. 768-771, 1997.

BERARDINI, Tanya Z. *et al*. Functional annotation of the Arabidopsis genome using controlled vocabularies. **Plant physiology**, v. 135, n. 2, p. 745-755, 2004.

BOYLE, Godfrey. Renewable energy: power for a sustainable future. **Oxford University Press**, 1996.

BRUNET, Marie A.; ROUCOU, Xavier. Mass spectrometry-based proteomics analyses using the OpenProt database to unveil novel proteins translated from non-canonical open reading frames. **JoVE (Journal of Visualized Experiments)**, n. 146, p. e59589, 2019.

BRUNET, Marie A.; LEBLANC, Sebastien; ROUCOU, Xavier. Reconsidering proteomic diversity with functional investigation of small ORFs and alternative ORFs. **Experimental cell research**, p. 112057, 2020.

CALVIELLO, Lorenzo; OHLER, Uwe. Beyond read-counts: Ribo-seq data analysis to understand the functions of the transcriptome. **Trends in Genetics**, v. 33, n. 10, p. 728-744, 2017.

CARLSON, Marc *et al*. GO. db: A set of annotation maps describing the entire Gene Ontology. **R package version**, v. 3, n. 0, p. 10.18129, 2017.

CAVALCANTE FILHO, Pedro Gilberto; BUAINAIN, Antônio Márcio; DE SOUZA BENATTI, Gabriela Solidario. A cadeia produtiva agroindustrial do biodiesel no Brasil: um estudo sobre sua estrutura e caracterização. **DRd-Desenvolvimento Regional em debate**, v. 9, p. 772-799, 2019.

CHENG, Haoyu *et al*. Small open reading frames: current prediction techniques and future prospect. **Current Protein and Peptide Science**, v. 12, n. 6, p. 503-507, 2011.

CHUGUNOVA, Anastasia *et al*. Mining for small translated ORFs. **Journal of proteome research**, v. 17, n. 1, p. 1-11, 2018.

COMBIER, Jean-philippe; LAURES-SERGUES, Dominique; BECARD, Guillaume. Use of micropeptides for promoting plant growth. **U.S. Patent** n. 10,563,214, 18 fev. 2020.

DA SILVA FONSECA, Nayanna Brunna; SOTO-BLANCO, Benito. Toxicidade da ricina

presente nas sementes de mamona. **Semina: Ciências Agrárias**, v. 35, n. 3, p. 1415-1424, 2014.

DAVIS, Melissa J.; SEHGAL, Muhammad Shoaib B.; RAGAN, Mark A. Automatic, context-specific generation of Gene Ontology slims. **BMC bioinformatics**, v. 11, n. 1, p. 1-13, 2010.

DE ALBUQUERQUE, F. A. *et al*. Crescimento e desenvolvimento do Pinhão manso: 1º ano agrícola. **Embrapa Algodão-Documentos (INFOTECA-E)**, 2008.

DE ARRUDA, FRANCINEUMA PONCIANO *et al*. Cultivo de pinhão manso (Jatropha curca L.) como alternativa para o semi-árido nordestino. **Revista brasileira de oleaginosas e fibrosas**, v. 8, n. 1, 2004.

DE CONINCK, Barbara; DE SMET, Ive. Plant peptides–taking them to the next level. **Journal of Experimental Botany**, v. 67, n. 16, p. 4791-4795, 2016.

DE MELO, Danielle Brandão *et al*. Estado da Arte do Cultivo e Produção de Pinhão-Manso para Biodiesel: Uma Revisão/State of the Art of Culture and Pinhão-Manso Production for Biodiesel: A Review. **Saúde em Foco**, v. 6, n. 2, p. 29-39, 2019.

DELCOURT, Vivian *et al*. Small proteins encoded by unannotated ORFs are rising stars of the proteome, confirming shortcomings in genome annotations and current vision of an mRNA. **Proteomics**, v. 18, n. 10, p. 1700058, 2018.

ERHARD, Florian *et al*. Improved Ribo-seq enables identification of cryptic translation events. **Nature methods**, v. 15, n. 5, p. 363-366, 2018.

FRITH, Martin C. *et al*. The abundance of short proteins in the mammalian proteome. **PLoS genetics**, v. 2, n. 4, p. e52, 2006.

GALINDO, Máximo Ibo *et al*. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. **PLoS biology**, v. 5, n. 5, p. e106, 2007.

GENE ONTOLOGY CONSORTIUM. The Gene Ontology (GO) database and informatics resource. **Nucleic acids research**, v. 32, n. suppl_1, p. D258-D261, 2004.

GOLLO, SILVANA SAIONARA et al. Configuração da cadeia produtiva do biodiesel, a partir da matéria-prima soja, no Rio Grande do Sul/Brasil. In: **Embrapa Amazônia Oriental-Artigo em anais de congresso (ALICE)**. Brasília, DF: Sociedade Brasileira de Economia, Administração e Sociologia Rural, 2010.

HANADA, Kousuke *et al*. A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection. **Genome research**, v. 17, n. 5, p. 632-640, 2007.

HANADA, Kousuke *et al*. Small open reading frames associated with morphogenesis are hidden in plant genomes. **Proceedings of the National Academy of Sciences**, v. 110, n. 6, p. 2395-2400, 2013.

HE, Wei *et al*. Analysis of seed phorbol-ester and curcin content together with genetic

diversity in multiple provenances of Jatropha curcas L. from Madagascar and Mexico. **Plant physiology and biochemistry**, v. 49, n. 10, p. 1183-1190, 2011.

HELLER, Joachim. Physic nut, Jatropha curcas L. **Bioversity international**, 1996.

KASTENMAYER, James P. *et al*. Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. **Genome research**, v. 16, n. 3, p. 365-373, 2006.

KOHLHEPP, Gerd. Análise da situação da produção de etanol e biodiesel no Brasil. **Estudos avançados**, v. 24, n. 68, p. 223-253, 2010.

LADOUKAKIS, Emmanuel *et al*. Hundreds of putatively functional small open reading frames in Drosophila. **Genome biology**, v. 12, n. 11, p. 1-17, 2011.

LAVIOLA, B. G.; ALVES, A. A.; KOBAYASHI, A. K.; FORMIGHIERI, E. F. Pinhão manso na EMBRAPA Agroenergia. Brasília: **INFOTECA-E**, 2015. 7 p. (Comunicado técnico, n. 12).

LEÃO, Evelynne Urzêdo; ADORIAN, Gentil Cavalheiro. TENDÊNCIAS DOS ESTUDOS COM MATÉRIAS PRIMAS ALTERNATIVAS PARA PRODUÇÃO DE BIODIESEL NO BRASIL. **Revista Integralização Universitária**, n. 21, p. 145-157, 2019.

MA, Jiao *et al*. Improved identification and analysis of small open reading frame encoded polypeptides. **Analytical chemistry**, v. 88, n. 7, p. 3967-3975, 2016.

MAKAREWICH, Catherine A.; OLSON, Eric N. Mining for micropeptides. **Trends in cell biology**, v. 27, n. 9, p. 685-696, 2017.

MAZUMDAR, Purabi et al. An update on biological advancement of Jatropha curcas L.: new insight and challenges. **Renewable and Sustainable Energy Reviews**, v. 91, p. 903-917, 2018

NEKRUTENKO, Anton; MAKOVA, Kateryna D.; LI, Wen-Hsiung. The KA/KS ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. **Genome research**, v. 12, n. 1, p. 198-202, 2002.

OCHMAN, Howard. Distinguishing the ORFs from the ELFs: short bacterial genes and the annotation of genomes. **TRENDS in Genetics**, v. 18, n. 7, p. 335-337, 2002.

OLEXIOUK, Volodimir; VAN CRIEKINGE, Wim; MENSCHAERT, Gerben. An update on sORFs. org: a repository of small ORFs identified by ribosome profiling. **Nucleic acids research**, v. 46, n. D1, p. D497-D502, 2018.

ORR, Mona Wu *et al*. Alternative ORFs and small ORFs: shedding light on the dark proteome. **Nucleic Acids Research**, v. 48, n. 3, p. 1029-1042, 2020.

PEREIRA, Illana Reis *et al*. Trends and gaps in the global scientific literature about Jatropha curcas L.(Euphorbiaceae), a tropical plant of economic importance. **Semina: Ciências Agrárias**, v. 39, n. 1, p. 7-17, 2018.

RAMOS, L. P. *et al*. Biodiesel: Um projeto de sustentabilidade econômica e sócio-ambiental para o Brasil. **Biotecnologia: Ciência e Desenvolvimento**, vol. 31, p. 28-37, 2003.

RICE, Peter; LONGDEN, Ian; BLEASBY, Alan. EMBOSS: the European molecular biology open software suite. **Trends in genetics**, v. 16, n. 6, p. 276-277, 2000.

RINALDI, R., GARCIA, C., MARCINIUK, L. L., ROSSI, A. V. SCHUCHARDT, U. Síntese de Biodiesel. Uma proposta Contextualizada de Experimento para Laboratório de Química Geral. **Química Nova**, v.30, n.5, p.1374-1380, 2007.

SIEBER, Patricia; PLATZER, Matthias; SCHUSTER, Stefan. The definition of open reading frame revisited. **Trends in Genetics**, v. 34, n. 3, p. 167-170, 2018.

VERLI, Hugo. Bioinformática: da biologia à flexibilidade molecular. 1ª ed. São Paulo: **Sociedade Brasileira de Bioquímica e Biologia Molecular**; 2014. 282 p.

YANG, Ziheng. PAML 4: phylogenetic analysis by maximum likelihood. **Molecular biology and evolution**, v. 24, n. 8, p. 1586-1591, 2007.

WICKHAM, Hadley *et al*. Elegant graphics for data analysis. **Media**, v. 35, n. 211, p. 10.1007, 2009.

WILKE, Claus O. (2020). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. **R package version 1.1.1**, 2020.

YESILYURT, Murat Kadir. The evaluation of a direct injection diesel engine operating with waste cooking oil biodiesel in point of the environmental and enviroeconomic aspects. **Energy Sources**, Part A: Recovery, Utilization, and Environmental Effects, v. 40, n. 6, p. 654-661, 2018.

ZHANG, Lin et al. Global analysis of gene expression profiles in physic nut (Jatropha curcas L.) seedlings exposed to salt stress. **Plos one**, v. 9, n. 5, p. e97878, 2014.

ZHU, Mengmeng; GRIBSKOV, Michael. MiPepid: MicroPeptide identification tool using machine learning. **BMC bioinformatics**, v. 20, n. 1, p. 1-11, 2019.