UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

NAYANE DE SOUZA

PEQUENAS ORFS SOB O OLHAR DA BIOINFORMÁTICA: UMA ANÁLISE CIENCIOMÉTRICA

DOIS VIZINHOS

2023

NAYANE DE SOUZA

PEQUENAS ORFS SOB O OLHAR DA BIOINFORMÁTICA: UMA ANÁLISE CIENCIOMÉTRICA

.

SMALL ORFS FROM THE BIOINFORMATICS PERSPECTIVE: A SCIENTOMETRIC ANALYSIS

Projeto de Trabalho de conclusão de curso de Especialização apresentado como requisito para obtenção do título de Especialização em Biologia Molecular – Habilitação Bioinformática da Universidade Tecnológica Federal do Paraná (UTFPR).
Orientadora: Nédia de Castilhos Ghisi

DOIS VIZINHOS

2023

NAYANE DE SOUZA

PEQUENAS ORFS SOB O OLHAR DA BIOINFORMÁTICA: UMA ANÁLISE CIENCIOMÉTRICA

Projeto de Trabalho de Conclusão de Curso de Especialização apresentado como requisito para obtenção do título de Especialista em Biologia Molecular – Habilitação Bioinformática da Universidade Tecnológica Federal do Paraná (UTFPR).

Data de aprovação: 27 de fevereiro de 2023.

_____

Henrique Moura Dias
Engenheiro Florestal
Universidade de São Paulo

_____

Thais Fernandes Mendonca Mota
Doutorado
Universidade Federal Tecnológica do Paraná – Dois Vizinhos

DOIS VIZINHOS

2023

Dedico este trabalho à minha família e amigos por todo o apoio.

AGRADECIMENTOS

Agradeço primeiramente a toda minha família e amigos por acreditarem em mim e me apoiarem durante toda minha vida.

Deixo um agradecimento especial a minha orientadora e professora Dra Nédia de Castilhos Ghisi, que despertou em mim o interesse por estudos sistemáticos da literatura e que compartilhou sua bagagem e conhecimento na disciplina que ministrou durante a especialização, esse aprendizado levarei para sempre em minha formação acadêmica e, com certeza, é fundamental para que todos nós desenvolvamos projetos de pesquisa com mais qualidade. Agradeço também pelo tempo dedicado e por acreditar em meu trabalho os quais foram fundamentais para que desenvolvesse esse estudo.

Agradeço também meu atual orientador de doutorado Dr. Alexandre Rossi Paschoal que foi fundamental para que eu tivesse coragem de iniciar esta etapa em minha carreira, pela dedicação de seu tempo que é bem escasso, pelo apoio e por acreditar em mim. Agradeço por compartilhar todo seu conhecimento e me apresentar ao mundo de pequenas ORFs, despertar em mim o interesse e curiosidade pela área e, principalmente, me acompanhar e orientar durante esse trajeto.

Também quero agradecer à Universidade Tecnológica Federal do Paraná de Dois Vizinhos e todos os professores, nós alunos acompanhamos durante esses dois anos a dedicação de toda a equipe para que o curso fosse de alta qualidade e realmente foi, agradeço por acreditarem na ciência, mesmo com todas as dificuldades, e dedicarem seu tempo na criação desta especialização, além da preocupação que tiveram com nosso aprendizado e opinião com relação ao curso.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Brasil (CAPES) que permite que tenhamos acesso aos periódicos gratuitamente.

“Science and everyday life cannot and should not be separated.”

Rosalind Franklin

RESUMO

As pequenas ORFs são regiões no genoma capazes de traduzir proteínas de até 100 aminoácidos e que por muito tempo foram consideradas lixo no processo de anotação. Entretanto, vários estudos reportaram funções fisiológicas e patológicas importantes em vários organismos e a área tem ganhado visibilidade e, consequentemente, aumento do número de dados. O uso de métodos de bioinformática possui um papel importante para identificação, predição e organização dessas pequenas ORFs e suas proteínas. Diante da relevância da análise cienciométrica para compreender o Estado da Arte e tendências em determinada área e a inexistência de tais estudos sobre pequenas ORFs e pequenas proteínas no contexto da bioinformática, realizei uma análise qualitativa e quantitativa sobre termos mais comuns, revistas de maior visibilidade, autores que mais publicam e correlações entre eles, instituições e países que se destacam e contribuem entre si. Incialmente foi realizada uma busca em duas bases de dados, Scopus e Web of Science, combinando termos relacionados a pequenas ORFs, pequenas proteínas e bioinformática e, de acordo com critérios de exclusão e inclusão pré-determinados, selecionei 82 trabalhos. A partir desses dados, foi obtido o índice H 21 que sugere uma boa relevância considerando o tamanho do conjunto analisado. Dentre os termos com maior frequência destaca-se câncer, peptídeos antimicrobianos e, curiosamente, a classe de peixes *actinoperygil*, apontando possíveis estudos aplicados nessas áreas. Com relação à co-ocorrência de termos destaca-se RNAs não codificantes, em especial os longos, corroborando com estudos anteriores que demonstram sua importância em diversas patologias. Por fim, as análises de autores, instituições e países mostram que este campo de estudo ainda é pouco explorado e restrito a poucas instituições e países, principalmente a China e Estados Unidos. Portanto, o aumento no número de publicações desde 2016 combinado a poucos grupos contribuindo, indicam que há espaço para que novos grupos de pesquisa ao redor do mundo possam fazer parte desta comunidade trazendo novos problemas, ideias e hipóteses, e então impulsionando o crescimento da área.

Palavras-chave: sORF; smORF; pequenas proteínas; pequenos peptídeos; microproteínas; bioinformática; Ribo-seq; RNA não-codificante.

ABSTRACT

Small ORFs are regions in the genome capable of translating proteíns shorter than 100 amino acids and had been considered as trash for a long time during the annotation process. However, many studies have shown important physiological and pathological functions in several organisms and the field has been gaining visibility and, consequently, the increase of the number of data. Bioinformatics application plays an important role at the identification, prediction and organization of small ORFs and their proteins. Considering the significance of scientometrics analysis to comprehend the State of the Art and trends in a specific área, besides the lack of these studies about small ORFs and small proteins in Bioinformatics, I performed a qualitative and quantitative analysis about the most comon terms, higher visibility, authors who publish the most and correlations among them, and prominent educational affiliations and countries that contribute the most. First, I explored two databases, Scopus and Web of Science, combining and searching for terms related to small ORFs, small proteins and bioinformatics according to pre-established exclusion and inclusion criteria, 82 studies were selected for further evaluation. Bases on these data, the H-index 21 has been obtained which suggests a good relevance considering the size of this dataset, besides, among the most frequent terms the highlights are cancer, antimicrobial peptides and, curiously, the fish class *actinoperygil*, indicationg possible applied studies in this area. Concerning the terms co-ocurrence the ones that highlight are non-coding RNAs, especially the long ones, corroborating with previous studies which shows their importance in different pathologies. Finally, authors, affiliations and countries analysis indicate that this field of study is still restricted to few institutions and countries, specially the United States and China. Hence, the increase of publications since 2016 allied to few contributing groups, indicate that there is a gap so that new research teams around the world can become part of this community bringing new problems, ideas and hypothesis, and then leading to the progress in this field.

Palavras-chave: sORF; smORF; small proteins; small peptides; microproteins; bioinformatics; Ribo-seq; non-coding RNA.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

AltORFs	ORFs alternativas (do inglês *alternative ORFs*).
AMP	Peptídeos antimicrobianos (do *inglês antimicrobial peptides*).
CDS	Regiões de sequência de codificação, sequência de referência (do inglê *Coding sequence regions*)
circRNA	RNA circular (do inglês circular RNA)
dORF	ORF após a sequência de referência (do inglês *Downstream ORF*).
HPP	Projeto do Proteoma Humano (do inglês *Human Proteome Project*).
HUPO	Organização do Proteoma Humano (do inglês *Human Proteom Organization*).
lncRNA	RNA longo não-codificante (do inglês *long non-coding RNA*)
miRNA	Micro RNA
mRNA	RNA mensageiro (do inglês messenger RNA)
MS	Espectrometria de massas (do inglês *Mass spectrometry*)
ncRNA	RNA não codificante (do inglês *non-coding RNA*).
NGS	Sequenciamento de próxima geração (do inglês *Next Generatio Sequencing*).
ORF	fase de leitura aberta (do inglês *Open reading frames*)
pri-miRNA	Micro RNA primário (do inglês *primary micro RNA*)
Ribo-Seq	Sequenciamento de fragmentos protegidos pelo ribossomo.
RPF	Fragmentos protegidos pelo ribossomo (do inglês, *Ribosome Protecte Fragments*)
SEP	pequenos peptídeos codificados,ou microproteínas (do inglês *Sma Encoded Peptides*)
sORF	Pequena ORF (do inglês *small* ORF)
uORF	ORF antes da sequência de referência (do inglês *Upstream ORF*)

SUMÁRIO

1 INTRODUÇÃO

As proteínas desempenham diversos papéis nos organismos e a identificação por métodos computacionais de sequências codificadoras de proteínas (CDS, *coding sequence regions*) geralmente tem como um dos critérios o tamanho mínimo de 100 aminoácidos (ICHIHARA; NAKAYAMA; MATSUMOTO, 2022). As fases de leitura aberta, mais conhecidas como ORFs (*Open Reading Frames*), são sequências de DNA no genoma entre um códon de início e um códon de parada que possivelmente serão traduzidas e geralmente somente as mais longas, mais de 300 nucleotídeos, são consideradas como CDSs durante a anotação (DINGER *et al.*, 2008). Entretanto, o desenvolvimento de novas tecnologias para sequenciamento de próxima geração (NGS, *Next Generation Sequencing*) associados à ferramentas de bioinformática possibilitaram a identificação de regiões codificadoras em locais do genoma até então assumidos como não-codificadores como as regiões intergênicas, DNA intrônico, RNAs não codificantes (ncRNAs, *non-coding RNAs*) e pseudogenes (ONG, Sheue Ni *et al.*, 2022).

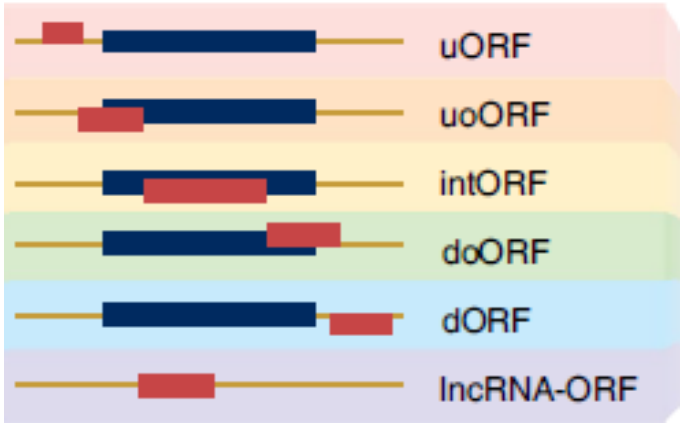
Um tipo de ORF que têm ganhado atenção é a pequena ORF, conhecida como sORF ou smORF, que são regiões que codificam microproteínas (ou SEPs, *Small Encoded Peptides*) que tem menos de 100 aminoácidos. O número de SEPs produzidas a partir dessas sORFs é relativamente discreto, contudo, estudos moleculares e estruturais revelaram que essas proteínas são significativamente versáteis e, em certos casos, participam de funções celulares essenciais. Além disso, há peptídeos não funcionais de sORFs que exibem função importante no sentido evolutivo e encontram-se extremamente conservadas (ORR *et al.*, 2020). Estudos nesta área têm aumentado consideravelmente nos últimos cinco anos, devido às recentes descobertas de importantes peptídeos oriundos dessas regiões (GUERRA-ALMEIDA; NUNES-DA-FONSECA, 2020).

Os avanços recentes das abordagens preditivas em bioinformática permitem um avanço significativo na identificação de milhares de prováveis sORFs, suas funções e organização dessas informações em bancos de dados.

1.1 Pequenas ORFs

As sORFs são regiões que codificam peptídeos com menos de 100 aminoácidos que podem ter codon de início AUG ou sinônimos como GUG, UUG, AUU, AUA, CUG, GUG e UUG. Essas sORFs podem estar localizadas em diversas regiões do genoma e as técnicas de perfil ribossômico (Ribo Seq*, Ribosome Sequencing*) e Espectrometria de Massas (MS, *Mass Spectrometry*) são grandes aliadas na identificação dessas pequenas ORFs. As sORFs podem estar antes (*upstream ORFs*) ou depois (*downstream ORFs*) das CDSs ou ainda sobrepondo-as (*overlapping ORFs*), elas também são encontradas no meio dessas CDSs (*intergenic ORFs*) ou mesmo em ncRNAs como lncRNAs e RNA circulares (COUSO; PATRAQUIM, 2017).

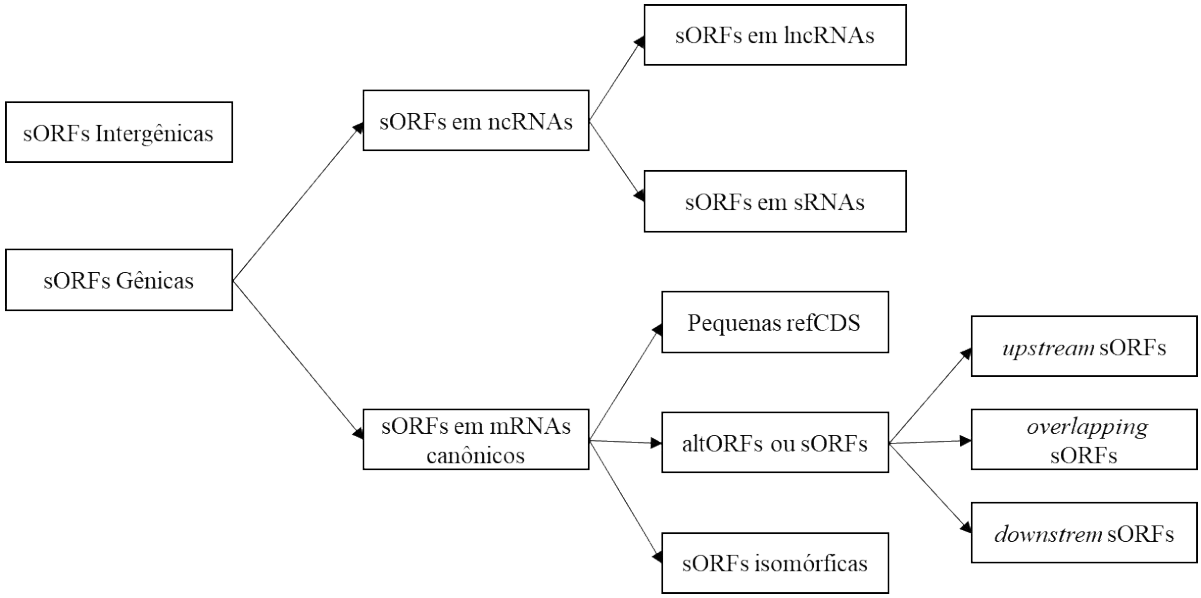
Figura 1 Os diversos tipos de pequenas ORFs de acordo com sua localização, em azul as CDSs e em vermelho as sORFs. uORF = upstream ORF, uoORF = upstream overlapping ORF, intORF = intergenic ORF, doORF = downstream overlapping ORF, dORF = downstream ORF e lncRNA- long non-coding RNA ORF.



Fonte: Adaptado de Couso e colaboradores (COUSO; PATRAQUIM, 2017).

Uma das formas de classificação é de acordo com seu perfil de transcrição, que primeiramente são divididas em pouco ou não expressas (intergênicas) e expressas (gênicas). O grupo de ORFs gênicas pode ser subdividido em localizadas em ncRNAs ou em mRNAs (*messenger RNA*) canônicos e, por fim, as sORFs em ncRNAs são subdivididas em diversas classes de acordo com o tipo do RNA não codificante (GUERRA-ALMEIDA, TSCHOEKE, NUNES-DA-FONSECA, 2021) (Figura 2).

Figura 2: Classificação de sORFs de acordo com seu perfil transcricional



Fonte: Adaptado de Guerra-Almeida e colaboradores (GUERRA-ALMEIDA; TSCHOEKE; NUNES-DA-FONSECA, 2021).

As sORFs intergênicas geralmente não são transcritas ou funcionais e estão localizadas entre uma sequência ATG e um códon de parada, alguns estudos já reportaram expressão de sORFs nessa região em *Salmonella Typhimurium* (CHOI et al, 2021) e sob condições de estresse em *E.coli*, neste caso indicando possivelmente expressão dependente do ambiente (HÜCKER et al, 2017). Uma das estratégias para encontrar novas sORFs intergênicas é incluindo a detecção de sORFs utilizando ferramentas de bioinformática, como o getORF da EMBOSS (RICE; LONGDEN; BLEASBY, 2000), seguido de filtros de conservação evolutiva e comparação com bancos de dados para verificar se as sequências são expressas (LADOUKAKIS et al., 2011).

Os ncRNAs possuem um papel importante em diversos organismos, em plantas atuam em várias funções vitais que englobam crescimento, desenvolvimento, e respostas a estresse abiótico em níveis transcricionais e pós-transcricionais (D'ARIO; GRIFFITHS-JONES; KIM, 2017). Os lncRNAs são moléculas de RNA com mais de 200 nucleotídeos que, a princípio, não codificam proteínas e vários deles já foram descobertos com importantes funções fisiológicas, como os *housekeeping* lncRNAs (GUERRA-ALMEIDA, TSCHOEKE, NUNES-DA-FONSECA, 2021; COUSO; PATRAQUIM, 2017). O padrão de aminoácidos dessas sequências é diferente das proteínas canônicas e não é aleatório, sugerindo um

novo padrão biológico, além disso podem não ter pontos de poliadenilação, sequências consenso Kozak e caps 5' (COUSO; PATRAQUIM, 2017). Essas características são importantes para uso em abordagens computacionais para detectar possíveis sORFs.

Em humanos os lncRNAs estão envolvidos, por exemplo, durante desenvolvimento e diferenciação celular (WU; YANG; CHEN, 2017) e quando desregulados estão relacionados com diversas doenças como o lncRNA DACORI que tem expressão reduzida em câncer de cólon (MORLANDO; FATICA, 2018), HOTAIR que em células de câncer de pulmão impacta proliferação, sobrevida, invasão, metástase e resistência a drogas (LOEWEN et al., 2014) e o LINCRNA-p21 que reduz a proliferação de células de câncer de próstata (WANG et al., 2017).

Os sRNAs (pequenos RNAS, do inglês *small RNAs*) possuem de 200 a 300 nucleotídeos e seus efeitos regulatórios já foram estudados em diversas espécies. A obtenção dessas moléculas, de modo experimental, depende exclusivamente do tipo de método utilizado para isolar o RNA visando não o perder, devido ao seu tamanho reduzido em relação aos demais tipos de RNA. Trabalhos sugerem seu efeito duplo, como por exemplo o SR1 da *B. subtilis* que age como RNA regulatório e como peptídeo estabilizando o RNA (GIMPEL; BRANTL et al, 2016). Portanto, uma das estratégias principais envolve a busca por sORF em pequenos RNAs regulatórios já conhecidos utilizando as ferramentas Expasy Translate Tool (ARTIMO et al, 2012) e posteriormente tBLASTn (ALTSCHUL et al, 1990).

As sORFs isomórficas são pequenas variantes de grandes ORFs que surgem principalmente por *splicing* alternativo, os peptídeos codificados por esse tipo de sORF tendem a seguir vias de atividade de seus precursores, bem como o padrão de uso de aminoácidos (GUERRA-ALMEIDA, TSCHOEKE, NUNES-DA-FONSECA, 2021; COUSO; PATRAQUIM, 2017).

As sORFs em regiões codificantes de proteínas de referência (refCDSs,) são as principais ORFs no mRNA que podem ser flanqueadas por sORFs (*upstream* e *downstream*). Suas características com relação ao padrão de aminoácidos e os exemplos já caracterizados indicam que essas SEPs possuem funções como reguladores de proteínas canônicas (COUSO; PATRAQUIM, 2017). As sORFs se diferem das chamadas ORFs alternativas (altORFs, *alternative ORFs*) no tamanho e códon de início. Enquanto as sORFs são regiões de 10 a 100 aminoácidos e possuem códon de início AUG ou não-AUG, as altORFs tem o tamanho mínimo de

30 aminoácidos sem limite superior e seu códon de início é AUG. (GUERRA-ALMEIDA, TSCHOEKE, NUNES-DA-FONSECA, 2021).

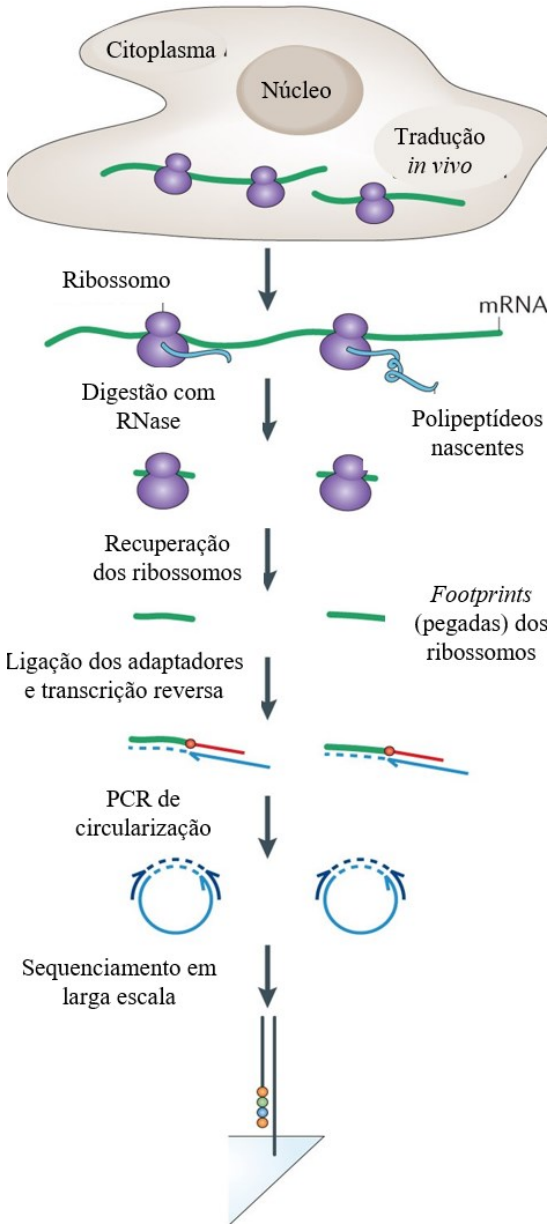
1.2 Técnicas experimentais para identificação sORFs e pequenas proteínas

Atualmente duas técnicas têm sido grandes aliadas para detecção de sORFs, o perfil ribossômico (Ribo-seq, do inglês *ribosome sequencing*) e espectrometria de massas (MS, *mass spectrometry*). Ainda há desafios mesmo com essas técnicas, abaixo descreverei cada uma delas, bem como suas principais vantagens e desvantagens.

1.2.1 Técnica de perfil ribossômico

A tecnologia de Ribo-seq começou a ser desenvolvida em 2009 (INGOLIA *et al.*, 2009) e, desde então, somos podemos detectar regiões ativamente traduzidas, capturando os fragmentos protegidos pelo ribossomo (RPF, *Ribosome Protected Fragments*), cada um dos RPFs possui cerca de 30 nucleotídeos (CHEN *et al.*, 2022) . Na figura 3 há um diagrama da técnica desde a lise celular para obtenção de ribossomos ligados ao mRNA até a obtenção dos dados.

Figura 3: Esquematização da técnica de perfil ribossômico (*ribosome profiling*). Os polissomos em tradução na célula são expostos e digeridos, entretanto alguns os fragmentos de RNA mensageiro (mRNA) estão protegidos pelo ribossomo, conhecido como *ribosome footprint* ou pegadas do ribossomo. Eles são recuperados e isolados de acordo com o tamanho e, finalmente, esses fragmentos são purificados, sequenciados e mapeados contra o genoma de referência.



Fonte: Adaptado de Ingolia, 2014 (INGOLIA, 2014).

O tratamento com inibidores de tradução como *harringtonine* ou *lactimidomycin* antes do sequenciamento impede que os ribossomos continuem o processo e, dessa forma, as regiões de início também podem ser identificadas, além da possibilidade de isolamento de ribossomos no processo de tradução baseado em afinidade e reduzindo a co-purificação de outras ribonucleoproteínas com os

ribossomos (LEONG *et al.*, 2022). As leituras provenientes do Ribo-seq também fornecem informações da posição e quantitativas, como onde a tradução ocorreu e qual a abundância do transcrito ao qual esses ribossomos estão ligados respectivamente. Ao normalizar essas leituras pela abundância do RNA do transcrito que eles ocupam, sua eficiência de tradução também pode ser estimada (KUTE *et al.*, 2021).

Os dados de Ribo-seq representam uma oportunidade e um desafio para os projetos de anotação. Isto porque com essa técnica podemos encontrar elementos traduzidos ainda não identificados e também validar anotações existentes. Entretanto, há dificuldades em desenvolver uma estrutura de anotação apropriada para esses conjuntos de dados, principalmente porque há questões importantes na interpretação biológica precisa de uma sORF traduzida, se ela é uma uORF, uoORF, dORF, doORF, lncOR ou ainda intORFs *(internal out-frame* ORF ou refCDS)? (MUDGE *et al.*, 2021)

Uma das preocupações é que Ribo-seq ORFs desafiam um princípio-chave na anotação de genes e proteínas onde “conservação = função”. A maioria das Ribo-seq ORFs não exibem grande conservação dos aminoácidos. Além disso, projetos para anotações de referência tem historicamente evitado a anotação de proteínas linhagem especifica “*de novo*” a não ser que exista evidência experimental clara fornecida. Além disso, Ribo-seq ORFs também desafiam a hipótese que “tradução = produção de proteína”. A técnica experimental em si não consegue distinguir uma ORF que produz uma proteína genuína de outra cujo produto da tradução é rapidamente degradado, somente monitora o evento da síntese e não o que vem posteriormente. Isto permite uma ampla variedade de explicações alternativas para os mecanismos biológicos da função Ribo-seq ORFs (MUDGE *et al.*, 2021).

Portanto, em frente a esses desafios, a espectrometria de massas pode ser uma grande aliada ao Ribo-seq para validar a existência da SEP.

1.2.2 Técnica de espectrometria de massas

A técnica de Ribo-seq fornece informações sobre a associação de ribossomos e RNA, enquanto que a espectrometria de massas (MS) identifica os

peptídeos resultantes da tradução. Essa identificação é importante porque apesar de ser esperado que uma SEP identificada no Ribo-seq seja realmente traduzida, a ausência de conservação dessas moléculas indica que a maioria não será funcional, ou ainda, serão degradadas rapidamente (KUTE *et al.*, 2021).

A MS é uma técnica adaptada para validar a expressão de peptídeos e proteínas e, além de validar a existência de SEPs, ela pode ser utilizada para entender sua função através da identificação de proteínas com as quais estão interagindo (interatoma) (FABRE; COMBIER; PLAZA, 2021).

Para a identificação de SEPs individuais podemos utilizar técnicas de superexpressão com marcadores repórter e técnicas baseadas em epítopos e de fluorescência. Porém, quando precisamos de detecção em larga escala as técnicas de MS são essenciais. Experimentos usando técnicas de MS mais comuns não são apropriados para detectar SEPs, isto porque elas poderiam ser mascaradas por produtos degradados por outras proteínas, ou não serem detectados devido a sua baixa abundância. Para contornar este problema poder ser usada outras técnicas de lise, digestão, cromatografia para separar peptídeos, separação de bandas de baixo peso molecular por SDS-PAGE, combinação com sequenciamento de RNA (RNA-seq), entre outras formas (KUTE *et al.*, 2021).

Um dos problemas atualmente no uso de dados provenientes de MS é referente à padronização, pois diferentes grupos usam diferentes metodologias e parâmetros. Historicamente, os projetos de anotação de referência sempre preferiram métodos de MS de alto rigor e a Organização do Proteoma Humano (HUPO, *Human Proteome Organization*) eo Projeto do Proteoma Humano (HPP, *Human Proteome Project*) tem como objetivo produzir uma anotação completa do proteoma humano e publicou regras para padronizar a evidência de MS para anotação dessas proteínas (ADHIKARI *et al.*, 2020; MUDGE *et al.*, 2021).

Dessa forma, a obtenção as técnicas de bioinformática capazes de prever determinadas SEPs afim de facilitar e limitar para análise de MS é essencial, além da importância de dados de MS padronizados e organizados de forma eficiente e completa são grandes aliados para que tenhamos acesso a dados completos, fáceis de encontrar, comparáveis e que possamos evoluir nessa área que ainda tem muito a explorada.

1.3 Pequenas ORFs em plantas, humanos e outros animais

As sORFs e as SEPs foram identificadas em diversos organismos, tais como bactérias, plantas e humanos, e desempenham diversos papéis fisiológicos e patológicos que estão sendo estudados para contribuir para saúde, agricultura e biotecnologia.

Os RNAs não codificantes lncRNA, miRNA e circRNA possuem diversas funções regulatórias, e além da ação dos transcritos eles podem conter sORFs que traduzem pequenas proteínas, as quais também são capazes de desempenhar importantes papéis nesses organismos. (DRAGOMIR *et al.*, 2020).

Um dos exemplos mais conhecidos de SEP foi descoberto em plantas e é codificado por dois transcritos primários de microRNA (pri-miRNA): pri-miR171b (*Medicago truncatula*) e o pri-miR165a (*Arabidopsis thaliana*) que são conhecidos pela produção dos peptídeos miPEP171b e miPEP165a, respectivamente, e são responsáveis pelo acúmulo de seus correspondentes miRNAs maduros, resultando na regulação negativa (*down regulation*) de genes alvo envolvidos no desenvolvimento da raiz (LAURESSERGUES *et al.*, 2015).

Os peptídeos antimicrobianos (AMP, *antimicrobial peptides*) são um grande grupo de pequenas proteínas secretadas por células procariotas e eucariotas como um mecanismo de defesa contra patógenos e para controlar a proliferação de competidores. As AMPs têm como alvo a membrana bacteriana e também alvos intracelulares como o ribossomo e, uma das AMPs mais estudadas é o peptídeo catiônico com 26 amino ácidos Mellitin (STEINBERG; KOCH, 2021). Ele é encontrado no veneno da abelha europeia *Apis mellifera* e, desde a primeira publicação em 1952, diversos efeitos antibacterianos já foram associados a ele, como por exemplo sinergismo com antibióticos convencionais para aumento de atividade (MEMARIANI *et al.*, 2019).

As SEPs também podem ter funções diferentes de seus ncRNAs. O lncRNA HOXB-AS3 é um dos lncRNAs anti-senso que sobrepõe o cluster em humano de genes de fatores de transcrição B(HOXB). O HOXB-AS3 possui um papel oncogênico em leucemia mieloide e sua alta expressão está associada com mau prognóstico (HUANG *et al.*, 2019). Entretanto, o peptídeo traduzido por ele desempenha uma função supressora tumoral em câncer colorretal e a perda desta

SEP é um evento oncogênico crítico que leva à reprogramação metabólica neste tipo de neoplasia (HUANG *et al.*, 2017).

Em plantas há várias SEPs já identificados no crescimento, morfologia e estresse abiótico (Tabela 1). Os SEPs proventientes de microRNAS (miPEPs) vêm sendo estudados com relação ao crescimento de plantas e alguns miPEPs já possui patente e está sendo pesquisado para promoção do crescimento de plantas (COMBIER, J.; LAURES-SERGUES, D.; BECARD, 2020). Este tipo de pesquisa pode impactar positivamente o cultivo de plantas de importância comercial, tanto para consumo interno quanto exportação.

Quadro 1: sORFs já identificadas em plantas e suas funções (*Predicted sORFs).

Função	Gene/peptídeo	Aminoácidos	Organismo	Referência
crescimento de plantas	POLARIS	36	Arabidopsis	(CASSON *et al.*, 2002)
	EPF1	104	Arabidopsis	(HARA *et al.*, 2007)
	Zm401	89	Milho	(WANG *et al.*, 2022)
	Zm908p11	97	Milho	(DONG *et al.*, 2013)
	TDIF	12	Zinnia, Arabidopsis	(FUKUDA, 2016)
	CLE	40-90	Arabidopsis	(FIUME; FLETCHER, 2012)
	Pri-miR171b	20	Alfalfa	(LAURESSERGUES *et al.*, 2015)
Morfologia	ROT4	53	Arabidopsis	(NARITA *et al.*, 2004)
	Brkl	84	Milho	(DJAKOVIC *et al.*, 2006)
	ENOD40 - A,B	12,24	Soja	(ROHRIG *et al.*, 2002)
	49 sORFs *	--	Arabidopsis	(HANADA *et al.*, 2013)
Estresse abiótico	KOD	25	Arabidopsis	(BLANVILLAIN *et al.*, 2011)
	OSIP108	10	Arabidopsis	(DE CONINCK *et al.*, 2013)
	FIS1, FIS2, FIS3	70-80	Soja	(NANJO *et al.*, 2011)
	36sORFs*	--	Arroz	(BASHIR *et al.*, 2014)
	CEP family	--	Arabidopsis	(TABATA *et al.*, 2014)
	CEPD1	99	Arabidopsis	(OHKUBO *et al.*, 2017)
	CEPD2	102	Arabidopsis	(OHKUBO *et al.*, 2017)
	sORF*	--	Arabidopsis	(HIGUCHI-TAKEUCHI *et al.*, 2020)

Fonte: Adaptado de (ONG, S N *et al.*, 2022)**.**

Esses são alguns exemplos de SEPs que mostram que esses peptídeos podem agir em diferentes organismos, em processos patológicos e fisiológicos e, com relação àqueles codificados por ncRNAs, podem desempenhar papel parecido com o ncRNA que o codifica ou totalmente oposto como no caso do HOXB-AS3.

1.4 Bioinformática, sORFs e pequenas proteínas

O estudo de sORFs é interdisciplinar e podemos falar em três principais abordagens complementares para detecção dessas moléculas, análises de bioinformática, estratégias de sequenciamento e análise proteômica como a espectrometria de massas (MS) (PEETERS; MENSCHAERT, 2020). As ferramentas de bioinformática utilizam diversas abordagens, como por exemplo o uso dos conceitos de conservação e similaridade com regiões bem anotadas para tentar encontrar sORFs (MARTINEZ et al., 2020). O sequenciamento com técnicas de RIBO - Seq ou sequenciamento de RNA (RNA-seq) fornecem evidência de transcrição e tradução (ZHU et al., 2018) e a MS prova a existência da proteína traduzida (MACKOWIAK et al., 2015).

O principal objetivo de métodos de bioinformática é distinguir elementos expressos e funcionais de ruídos aleatórios, para isso a conservação de sequência considerada um *hallmark* de sORFs funcionais (MAKAREWICH; OLSON, 2017), similaridade de sequências, e alinhamento com proteínas já conhecidas para obter potencial codificante e funções, são as métricas utilizadas para essas análises (PEETERS; MENSCHAERT, 2020).

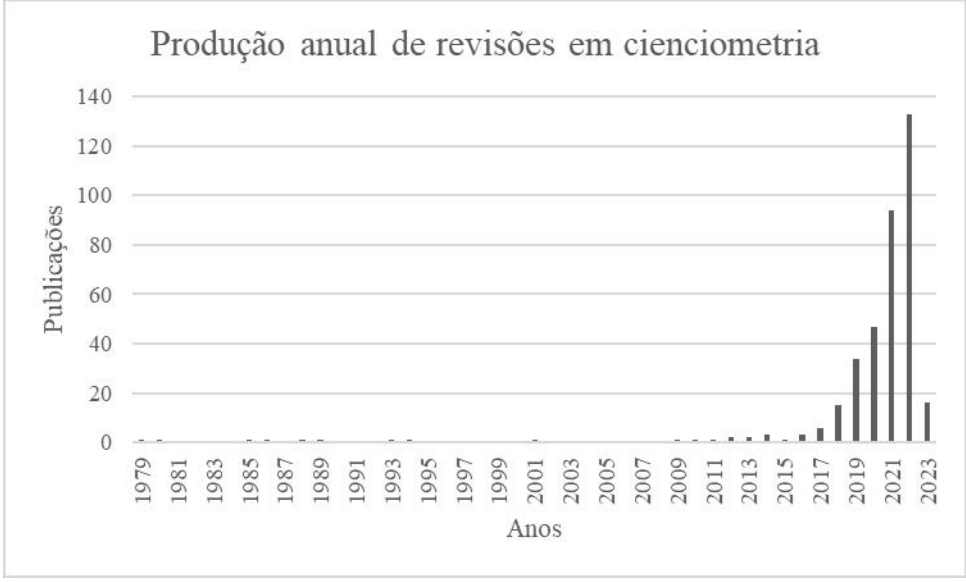
Em vista da grande quantidade de dados de sequenciamento publicados e importante papel dessas pequenas proteínas já demonstrados em trabalhos anteriores, a bioinformática tem uma função indispensável na identificação, seleção, organização em bases de dados e predição de sORFs e suas funções e, essas informações podem ser posteriormente usadas para validação experimental e diminuir custos com testes de muitas possíveis pequenas proteínas que podem excluídas em uma primeira avaliação em bioinformática.

1.5 Análise cienciométrica

O desenvolvimento de uma pesquisa envolve identificação de um problema e propostas de soluções. Entretanto, em vista da grande quantidade de trabalhos que são publicados anualmente pode ser difícil acompanhar como está a área de interesse, qual o estado da arte na pesquisa, quais perguntas ainda estão em aberto e quais instituições, autores e países estão mais envolvidos no assunto.

Artigos de revisão em cienciometria foram intensivamente discutidos entre 1970 e 1980 (BLÜMEL; SCHNIEDERMANN, 2020) e, embora o número de publicações tenha crescido desde de 2018 (Gráfico 1), ainda é relativamente baixo quando comparamos uma área específica e o número de publicações em cienciometria (Quadro 1).

Gráfico 1: Produção anual de artigos de revisão que possuem o termo *scientometric* no título.



Fonte: Dados obtidos no Web of Science em 01 de fevereiro de 2023.

No quadro 1 vemos uma combinação dos termos *scientometric** e *cancer* e vemos que ao usá-los como *title* e *topic* respectivamente, há uma grande diminuição no número de resultados, há 1551815 publicações em câncer e somente 57 com *scientometric** no título.

Quadro 2: Combinações de termos e número de resultados.

Title	Topic	Review	Resultados
----	Scientometric*	Todas as publicações	3517
----	Scientometric*	Sim	753
Scientometric*	----	Todas as publicações	1675

Scientometric*	----	Sim	399
Cancer	----	Todas as publicações	1551815
Cancer	----	Sim	126016
----	Cancer	Todas as publicações	3001841
----	Cancer	Sim	299599
Scientometric*	Cancer	Todas as publicações	57
Scientometric*	Cancer	Sim	11

Fonte: Dados obtidos no Web of Science em 01 de fevereiro de 2023.

A análise cienciométrica nos permite quantificar e analisar a produção científica disponível, sendo possível mapear autores, países, principais campos de pesquisa e quais os mais significativos (CHEN; SONG, 2019). De acordo com minha pesquisa antes da realização deste trabalho, esta é a primeira análise cienciométrica na área de pequenas ORFs e pequenas proteínas em bioinformática.

:

2 OBJETIVOS

2.1 Objetivo geral

Descrever o progresso científico na área de pequenas ORFs e pequenas proteínas associadas a estudos de bioinformática.

2.2 Objetivos específicos

- Identificar a base intelectual e de pesquisa sobre sORFs e pequenas proteínas na área de bioinformática;
- Identificar quais são os termos relacionados com a área e como estão conectados;
- Identificar os países e instituições mais influentes na área de pequenas ORFs e pequenas proteínas e como interagem entre si;
- Identificar os estudos e autores mais influentes, co-citações e colaborações.

3 METODOLOGIA

Para o desenvolvimento deste trabalho foi realizada uma busca sistemática por trabalhos em bioinformática sobre pequenas ORFs ou pequenas proteínas, refinamento e posterior análise em softwares para análise bibliométrica.

3.1 Busca sistemática por trabalhos sobre sORFs em bioinformática

A princípio, foi realizada uma busca nas bases Web of Science e Scopus para encontrar possíveis revisões já publicadas sobre pequenas ORFs e pequenas proteínas na bioinformática e, somente dois trabalhos foram encontrados os quais foram lidos cautelosamente e os principais termos utilizados para se referir ao tema foram usados para busca posterior (Quadro 1).

Quadro 3: Revisões encontrada sobre sORFs e pequenas proteínas.

Autores	Nome da publicação	Revista, ano, volume e páginas	DOI
Tharakan R, Sawa A.	Minireview: Novel Micropeptide Discovery by Proteomics and Deep Sequencing Methods	Front Genet. 2021 May 6;12:651485	10.3389/fgene.2021.651485
Pan J, Wang R, Shang F, Ma R, Rong Y, Zhang Y.	Functional Micropeptides Encoded by Long Non-Coding RNAs: A Comprehensive Review	Front Mol Biosci. 2022 Jun 13;9:817517.	10.3389/fmolb.2022.817517

Fonte: Autoria própria.

A partir dos termos já conhecidos relacionados às pequenas ORFs e pequenas proteínas, bem como aqueles encontrados nas revisões citadas, realizei uma seleção dessas palavras, as quais combinadas com booleanos e asteriscos restingira, e especificaram as buscas nas bases de dados (Quadro 2).

Quadro 4: Termos utilizados e combinações de caracteres booleanos para busca nas bases de dados SCOPUS e Web of Science.

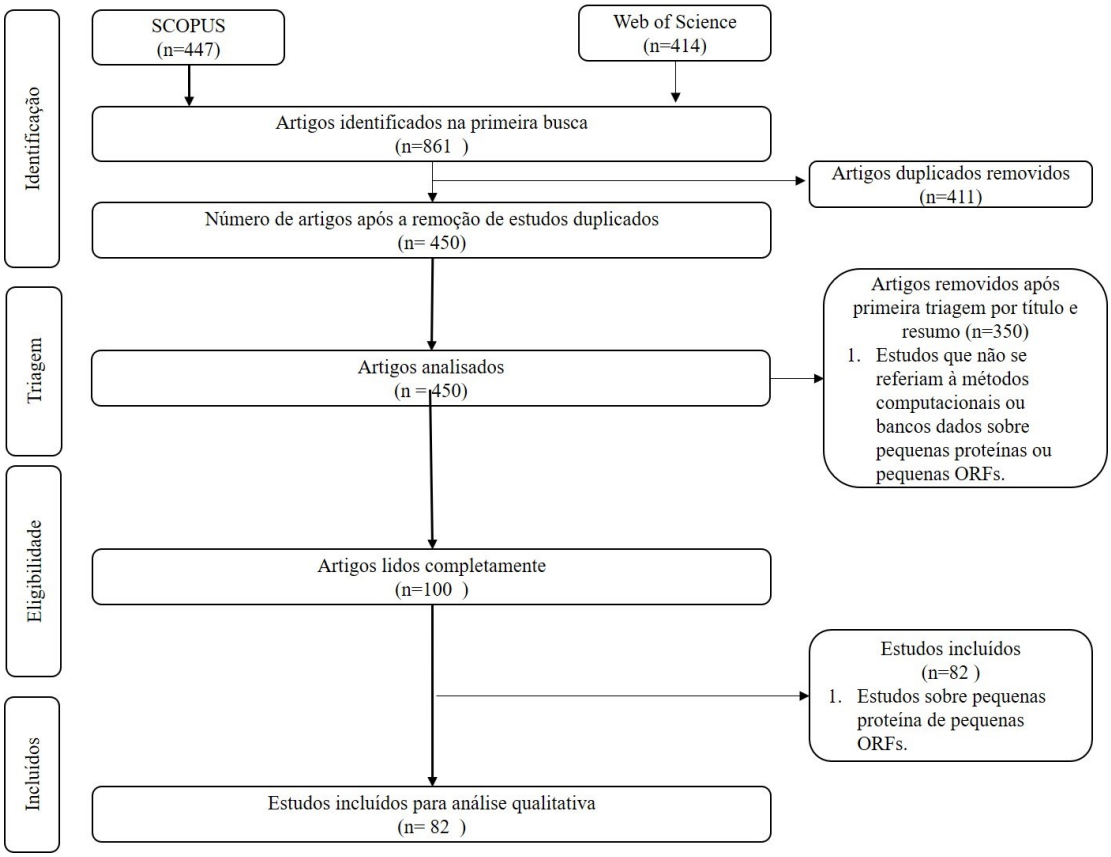
Bases de dados	Termo de busca
	Data da busca: 11 de dezembro de 2022

SCOPUS	TITLE-ABS-KEY ("sORF" OR "small ORF*" OR "smORF" OR "small peptide*" OR "small protein*" OR "short peptide*" OR "short protein*" OR "micropeptide*" OR "Small secreted peptide*" OR "small secreted protein*") AND TITLE ("in silico" OR tool* OR pipeline* OR workflow OR software OR algorithm* OR computational OR "machine learning"OR "database*" OR repository)
Web of Science	TS= ("sORF" OR "small ORF*" OR "smORF" OR "small peptide*" OR "small protein*" OR "short peptide*" OR "short protein*" OR "micropeptide*" OR "Small secreted peptide*" OR "small secreted protein*") AND TI = ("in silico" OR tool* OR pipeline* OR workflow OR software OR algorithm* OR computational OR "machine learning"OR "database*" OR repository)

Fonte: Autoria própria.

A seleção dos artigos baseada em critérios de exclusão e inclusão está representada no fluxograma prisma (Figura 4).

Figura 4: Fluxograma Prisma com os resultados obtidos em cada banco de dados, critérios de exclusão e inclusão.



Fonte: Autoria própria.

3.2 Indicadores bibliométricos

Para a avaliação bibliométrica foram utilizados os 82 artigos obtidos após leitura completa (Figura 4). Nesta análise, foi utilizado o pacote Bibliometrix no software RStudio 2022.12.0+353 (R versão 4.2.1) e interface web Biblioshiny (ARIA; CUCCURULLO, 2017).

O CteSpace foi utilizado para criar infográficos, valores de centralidade e nos ajudar a visualizar quais tendências de pesquisa, colaborações por países, instituições e referências. Cada círculo (*node*) no mapa representa um item e aqueles com maiores centralidade possuem um anel roxo ao seu redor, ela ainda varia de 0 a 1 e representa a influência de um país em determinada área. Além disso, as linhas que ligam esses círculos mostram as correlações entre eles. O tamanho da fonte também varia dependendo do número de publicações de cada item. Os aglomerados (*clusters*) formados indicam a semelhanças entre os objetos que os círculos representam (CHEN, 2006).

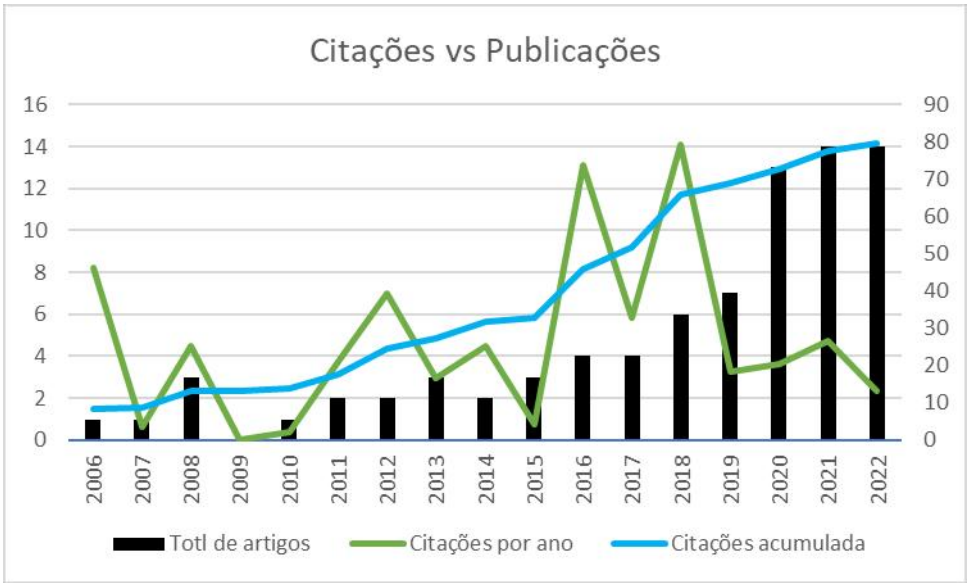
4 RESULTADOS E DISCUSSÃO

O conjunto de dados foi obtido em 11 de dezembro de 2022 com um total de 450 estudos, restando 82 após seleção. Este grupo possui índice H 21, ou seja, pelo menos 21 artigos foram citados por pelo menos 21 vezes cada, esse valor indica relevância desses estudos e contribuição na pesquisa, já que estamos falando de um número relativamente baixo de publicações analisadas (BURRELL, 2007)

As publicações sobre sORFs e pequenas proteínas em bioinformática que incluem bases de dados, análises *in silico*, ferramentas para identificação dessas regiões, estudos espécie-específica para identificação de regiões codificadoras ou predição de funcionalidade por análise computacional, cresceram desde a publicação de um banco de dados em 2006 *The arabidopsis unannotated secreted peptide database, a resource for plant peptidomics* que traz peptídeos secretados por essa planta entre 25 e 250 aa. Neste trabalho os autores estavam interessados em mapear peptídeos envolvidos na sinalização celular em *A. thaliana* e para isso eles utilizaram dados disponíveis de proteínas em *A. thaliana*, eliminaram características que impediriam que a proteína codificada pela ORF ficasse associada à membrana, afinal eles estavam procurando por moléculas envolvidas com sinalização celular, e finalmente usaram dados de *microarray* e PCR para avaliar a expressão dessas ORFs (LEASE; WALKER, 2006).

Entretanto, esse estudo foi publicado antes de um dos maiores aliados para pesquisa de sORFs, a técnica Ribo-Seq publicada em 2009 (INGOLIA *et al.*, 2009). Além desta metodologia, o uso mais acessível de sequenciamento de nova geração (NGS, *Next Generation Sequencing*) colaborou para um aumento no número de dados e, consequentemente métodos computacionais (Figura 5)

Gráfico 2 Número de publicações anuais e citações por ano e acumulada entre 2006 e 2022.



Fonte: Dados obtidos do Web of Science e Scopus.

O número de citações variou ao longo dos anos, mas até 2018 acompanhou as variações no número de publicações anual. Entretanto, a partir de 2020, apesar do aumento de publicações vemos um comportamento oposto com relação às citações. Ao olharmos para o número acumulado (linha azul), observamos que em 2016 teve uma variação expressiva com relação às citações de 2014, sendo que 2015 representa um ano com baixo número de citações comparado ao ano anterior e, principalmente, posterior.

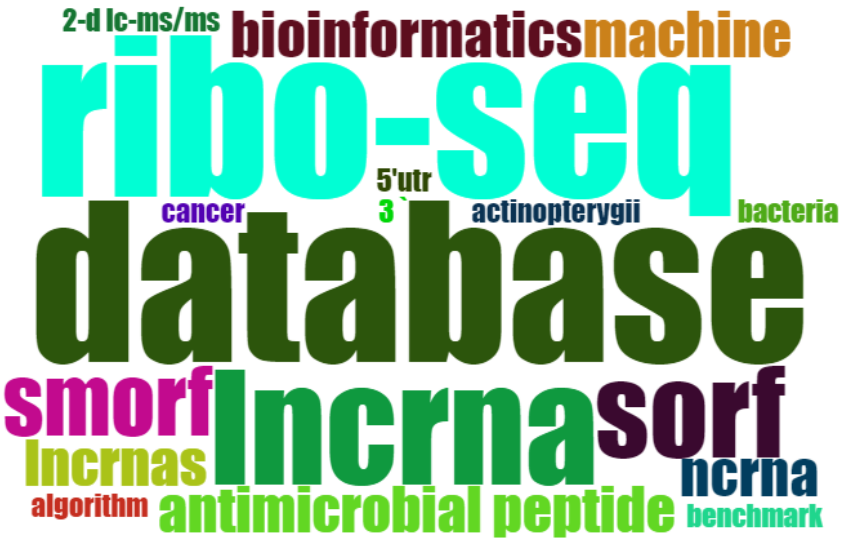
4.1 Nuvem de palavras e rede de co-ocorrência

Na nuvem de palavras obtida a partir dos trabalhos (Figura 5), vemos que os termos citados com maior frequência são *Database*, Ribo-seq, lncRNA, sORF, smORF, *bioinformatics* e *machine* sendo que esse último provavelmente se refere à *machine learning* (aprendizado de máquina). O maior número de frequência dessas palavras é esperado devido ao filtro utilizado, apesar da grande diferença entre *Database* e *machine* onde ambos os termos foram termos utilizados com o booleano "OR" nas strings. Esta discrepância entre *Database* e *machine* pode indicar que no contexto estudado os bancos de dados são mais relevantes do ponto de vista de número de artigos publicados.

Alguns termos como lc-ms/ms, Ribo-seq, cancer, *actinoperygil* (classe de peixes), *bacteria e antimicrobial peptide* nos mostra quais as técnicas mais citadas

que já eram esperadas, Ribo-seq e a espectrometria de massas. Os peptídeos antimicrobianos e bactérias são uma área de pesquisa importante para o tratamento de infecções multirresistentes, câncer que é uma das patologias mais pesquisadas e uma classe de peixes (*actinoperygii*) que chama a atenção, um dos estudos que relaciona essa classe de peixes com sORFs é uma anotação *de novo* de montagem do transcriptoma de *Apteronotus leptorhinchus* (da classe *actinoperigii*)que traz os RNAs expressos no tecido do sistema nervoso central (SALISBURY *et al.*, 2015).

Figura 5: Nuvem de palavras com maior frequência

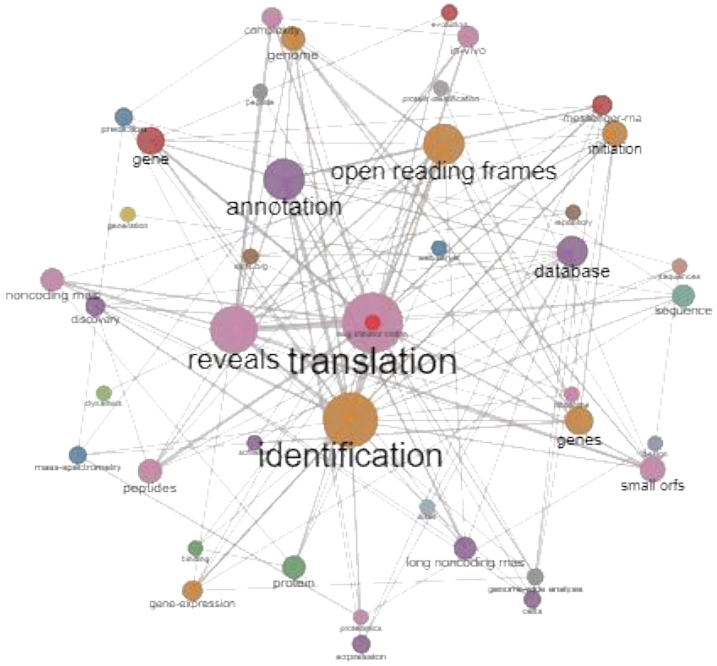


Fonte: Criado no Biblioshiny (ARIA; CUCCURULLO, 2017)**.**

A palavra *benchmark* também aparece nesta nuvem indicando que a comparação entre ferramentas e bancos de dados é um tema relevante neste tipo de estudo. Essas informações indicam quais as áreas mais pesquisadas e como podemos expandir as pesquisas analisando esses resultados.

A rede de palavras mais representativas obtidas nos bancos de dados representada na figura 6, onde cada círculo é proporcional à quantidade de publicações contendo a palavra-chave e a espessura das linhas que os conectam representa a quantidade de vezes que esses termos aparecem juntos em uma publicação. As palavras-chaves formam *clusters* agrupados por tema e representados por mesma cor dos círculos. Os maiores clusters nessa imagem são o rosa (*translation*, *reveals*, small ORFs, *non coding RNAs*, *peptides*, *in vivo*, *complexity*), laranja (*identification*, *open reading frames*, genes, *gene expression*, *genome*, *initiation*) e roxo (*annotation*, *database*, *long non coding RNAs*, *expression*). Esses resultados indicam um maior interesse no estudo de sORFs em RNAs não-codificantes, em especial os lncRNAs e a identificação de ORFs e sORF que regulam a expressão de genes.

Figura 6: Co-ocorrência de palavras-chave nos estudos selecionados.



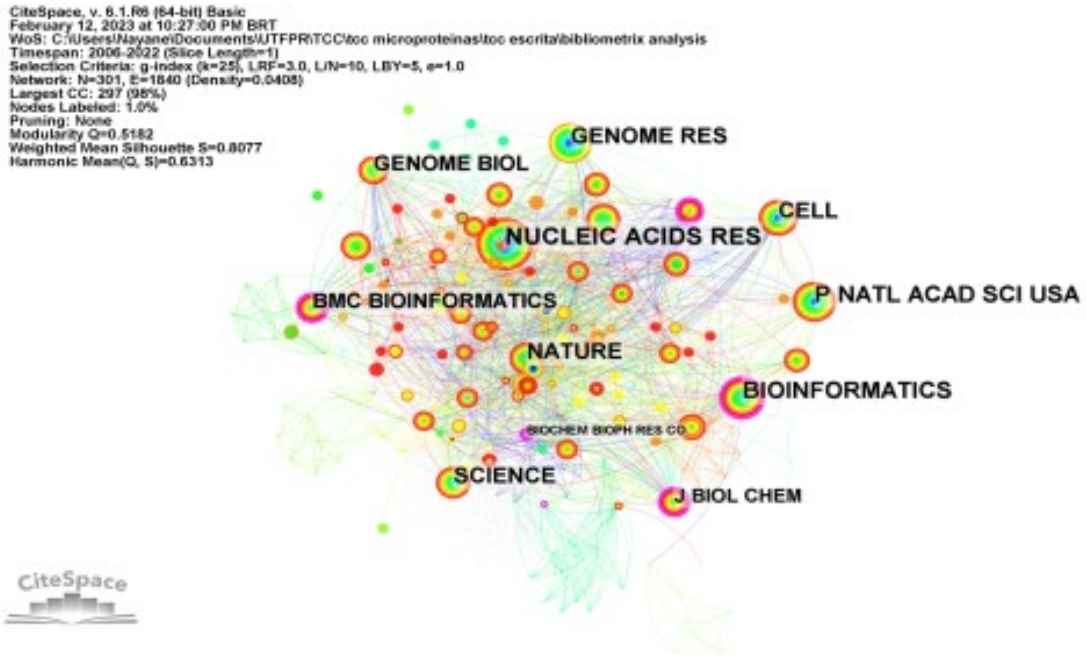
Fonte: Criado pela ferramenta Bibliometrix do R versão 4.2.1 (ARIA; CUCCURULLO, 2017).

A co-ocorrência de palavras nos indicou a importância dos RNAs não-codificantes, principalmente os lncRNAs, na pesquisa de sORFs. Isso faz sentido, já que cerca de 80% do genoma humano tem a capacidade de ser transcrito em ncRNAs. Os lncRNAs estão associados ao aparecimento de doenças, como câncer (LUO *et al.*, 2022) e os peptídeos de sORFs em lncRNAs representam atualmente uma nova fronteira nos estudos biomédicos focados em novos biomarcadores e alvos moleculares em câncer (GUERRA-ALMEIDA; TSCHOEKE; NUNES-DA-FONSECA, 2021)

4.2 Visibilidade e explosão de citações das revistas

As revistas científicas podem ser avaliadas por diversos parâmetros e o fator de impacto é um dos primeiros a ser pesquisado por grande parte dos cientistas. Contudo, elas possuem visibilidade diferentes dependendo do tema de pesquisa e esse é um fator importante quando decidimos onde publicar. Neste tópico de sORFs e pequenas proteínas relacionadas à bioinformática, as revistas que se destacam são, a *Nucleic Acids Research* que apresenta o maior número de citações (72), seguida da *Bioinformatics* (55), *P NatlAcad Science* (53), *Nature* (48), *Science* (46) e *Cell* (46). Elas estão representadas na figura 7 e, quanto maior o círculo e a fonte maiora visibilidade da revista.

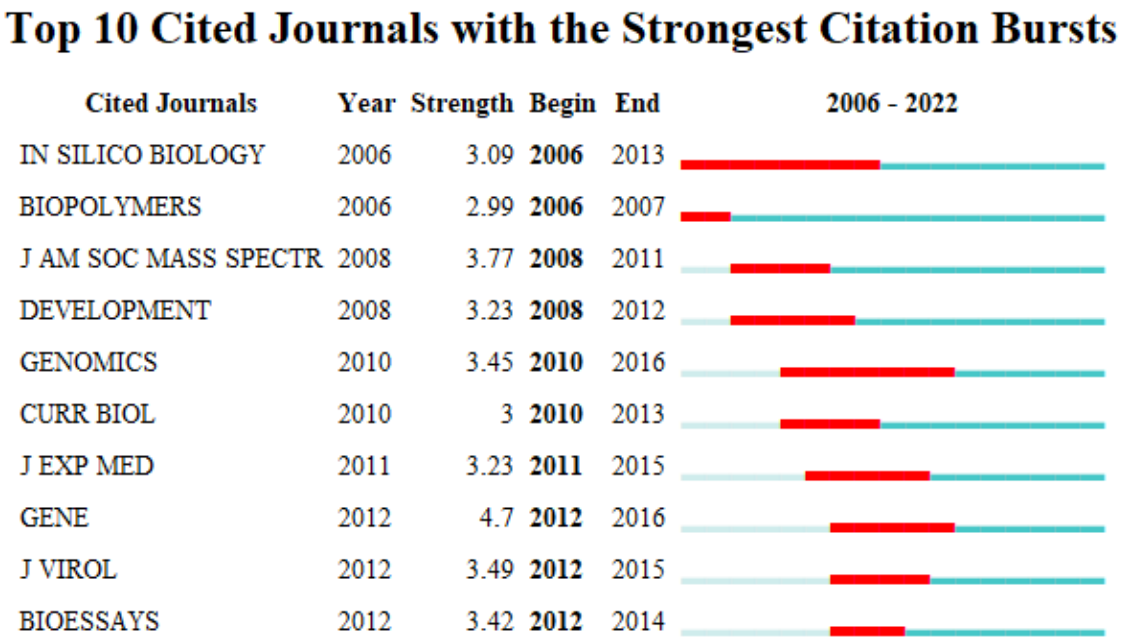
Figura 7: Rede de revistas e a visibilidade representada pelo tamanho dos círculos e fontes.



Fonte: Criado no CiteSpace (CHEN; DUBIN; KIM, 2014).

A explosão de citações (*Citation Burst*) representa o surgimento das citações e são indicadores de tendências, além de fornecer evidências que determinada publicação está associada com surgimento de trabalhos que atraíram muita atenção da comunidade científica (CHEN; DUBIN; KIM, 2014). A revista *Gene* possui a maior força de explosão (4.7), enquanto *Biopolymers* possui a menor (2.99) neste conjunto das 10 maiores explosões (Figura 8). No entanto, quando falamos em período que permaneceram nessa posição a revista *In Silico Biology* é a que durou mais tempo (7 anos) e a *Biopolymers* menos tempo (1 ano).

Figura 8: As 10 revistas com maiores explosões de citações.



Fonte: Criado no CiteSpace (CHEN; DUBIN; KIM, 2014).

As informações de visibilidade e explosão de citações são análises diferentes, sendo a primeira uma representação atual que nos ajuda a escolher revistas para publicar, e a segunda nos traz informação sobre o início de novas tendências na área e quanto tempo permaneceram, nos ajudando a compreender mais o campo de pesquisa e quais publicações representam importantes referências.

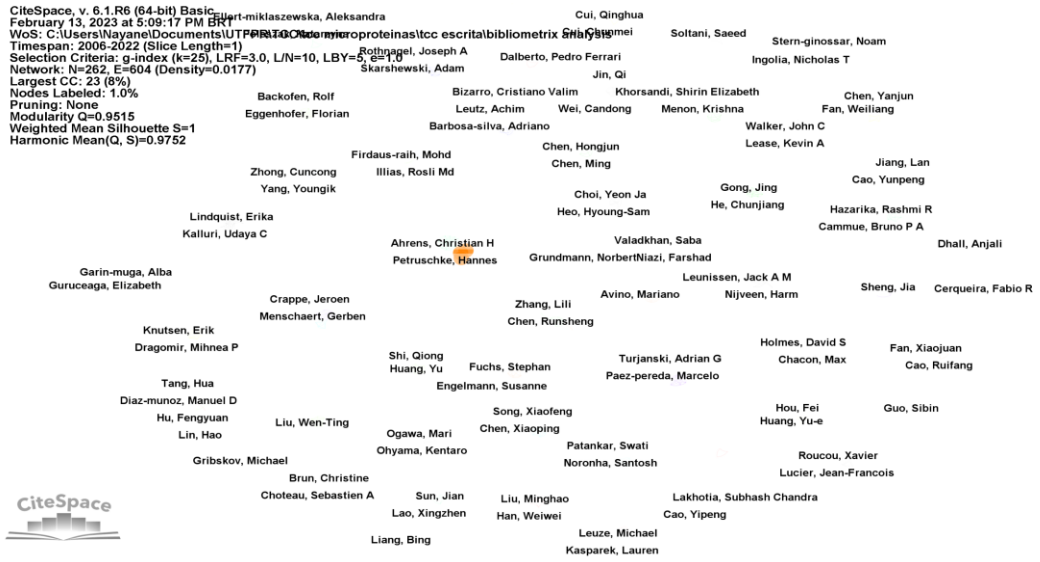
Dentre as seis revistas com maior visibilidade, as três primeiras neste critério possuem os fatores de impacto menor do grupo, *Nucleic Acids Research* com fator de impacto 19.16, seguida da *Bioinformatics* com 6.94, *Proceedings of the National Academy of Sciences* com 12.78, *Nature* 47.78, *Science* 41.84 e *Cell* 38.64. A revista *Bioinformatics* ganha uma posição interessante quando analisamos a visibilidade, já que entre as revistas com maiores fatores de impacto ela ocupa a posição 92ª.

4.3 Produção científica e colaboração entre autores

A colaborações entre autores é importante para responder problemas na área e diminuição custo em pesquisas. Além disso, essa troca de informações está

associada a estudos de maiores impactos e um número maior de publicações. A rede de autores relacionados ao tema deste trabalho (Figura 9) não tem uma centralidade tão alta (0,01 o máximo), indicando que esses autores não influenciam tanto outros trabalhos. Na figura 9 não é possível ver com clareza o nome dos autores e nem as correlações entre eles, isto porque ao realizar a análise eles ficaram muito distantes e, propositalmente, reduzi o tamanho para termos ideia dessa baixa correlação e baixíssima centralidade representada pelo ponto laranja.

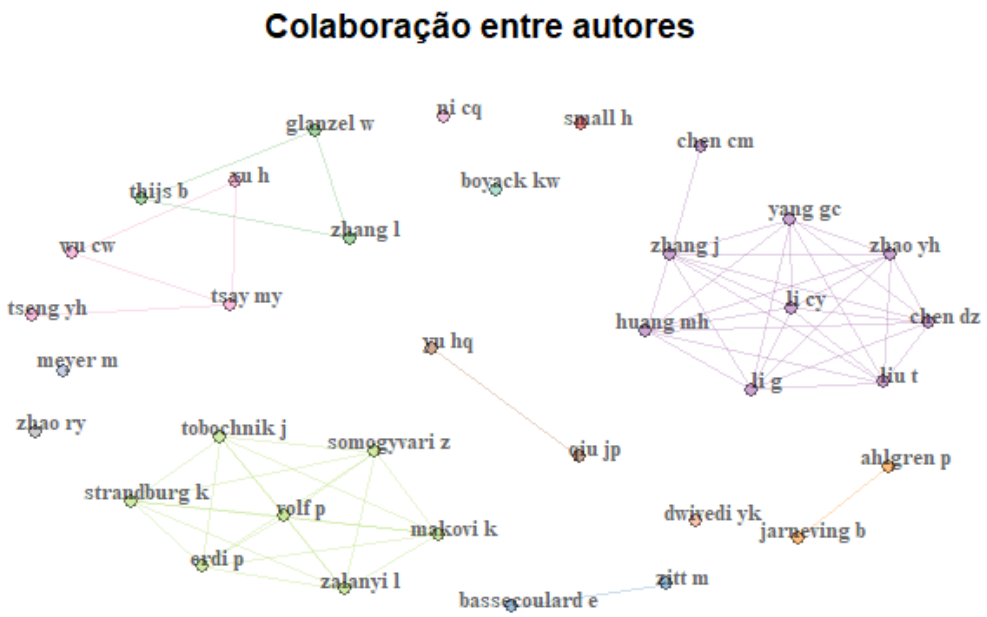
Figura 9: Rede de colaboração entre autores.



Fonte: Criado no CiteSpace (CHEN; DUBIN; KIM, 2014**).**

Além da centralidade, podemos olhar de fato como esses autores estão colaborando e o maior grupo possui 9 autores, seguido por outro com 7 autores, outro com 4, outro com 3, e 3 duplas, e 6 autores que não fazem colaboração (Figura 10).

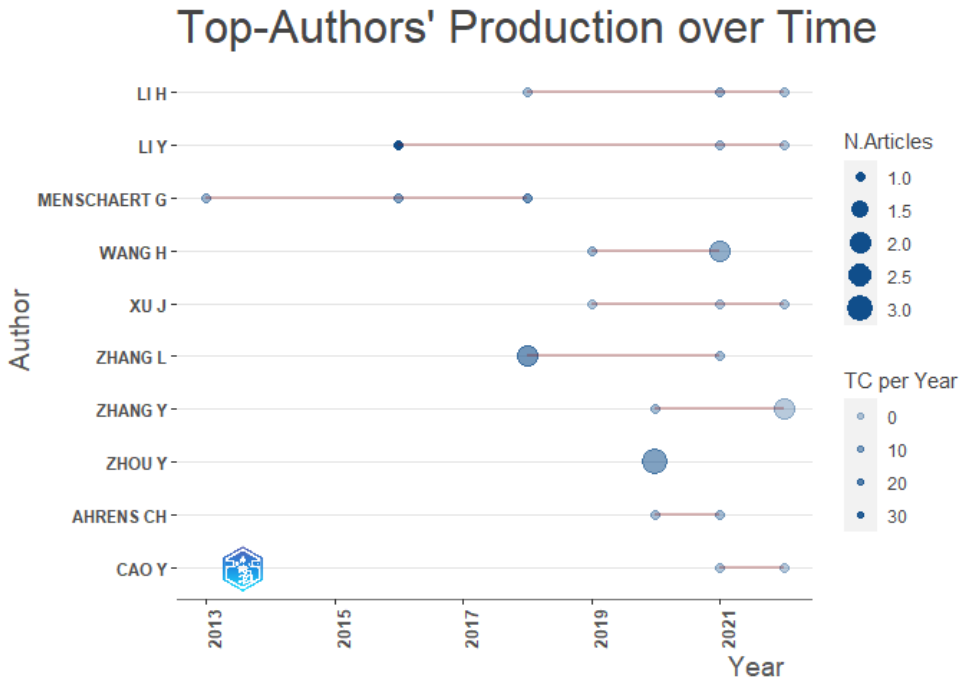
Figura 10: Rede de colaboração entre autores.



Fonte: Criado com a ferramenta Bibliometrix do R versão 4.2.1 (ARIA; CUCCURULLO, 2017)

Quando analisamos a produção científica dos autores (Figura 11), vemos que o máximo de publicações por autor na área é 3. Este é um número baixo, mas coerente com a baixa correlação e contribuição entre eles.

Figura 11: Autores com maior produção científica.



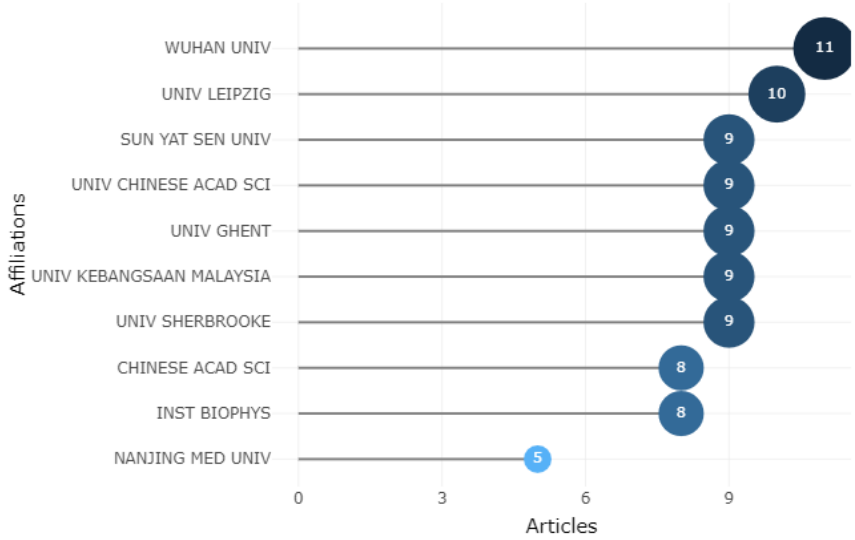
Fonte: Criado com a ferramenta Bibliometrix do R versão 4.2.1 (ARIA; CUCCURULLO, 2017).

A produção na área por autor, colaborações e correlações entre eles mostra que os trabalhos no geral têm sido realizados de maneira independente o que pode impactar negativamente a produção científica. Além disso, o tema abordado é interdisciplinar contendo pelo menos duas grandes áreas, Biologia e Informática, e esses resultados podem indicar também que esses campos de estudo não estão conversando afim de compreender de maneira mais ampla o problema de pesquisa que os permeia.

4.4 Contribuição científica de instituições e países

Com relação à distribuição de pesquisas nas instituições e no mundo, a *Wuhan University* na China possui o maior número de publicações (11), seguida pela *University of Leipzig* na Alemanha (10 publicações) e com 9 publicações temos a *Sun Yat Sem University* e *University of Chinese Academy of Sciences* na China, *University of Ghent* na Bélgica, *University of Kebangsaan* na Malásia e *University Sheerbrooke* no Canadá (Figura 12).

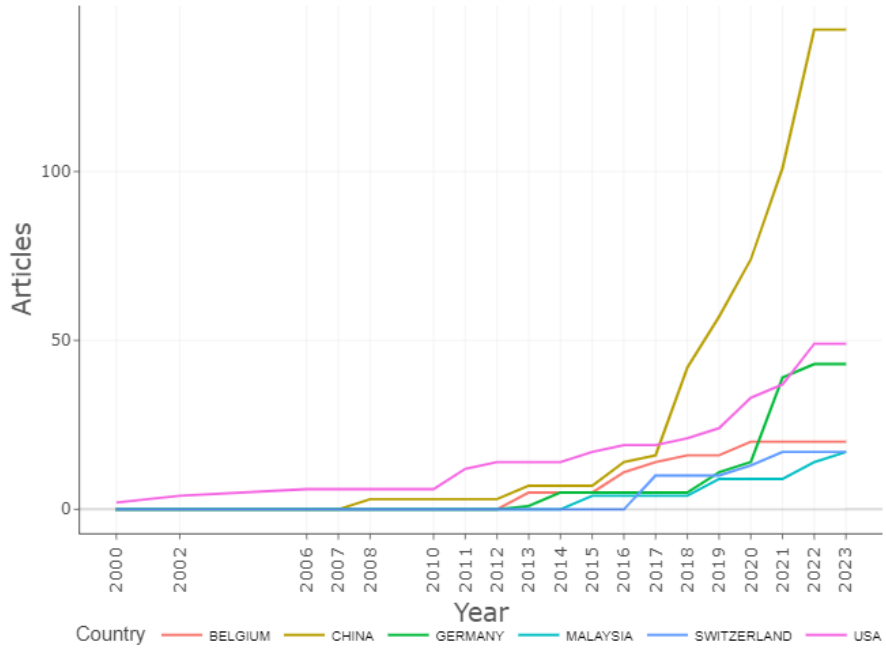
Figura 12: As 10 instituições com maior número de publicações na área.



Fonte: Criado na versão web do bibliometrix, Biblioshiny(ARIA; CUCCURULLO, 2017)

Com relação à produção por país, a China possui grande destaque (Figura 13) o que corrobora com as instituições com mais publicações, onde 5 de 10 são chinesas.

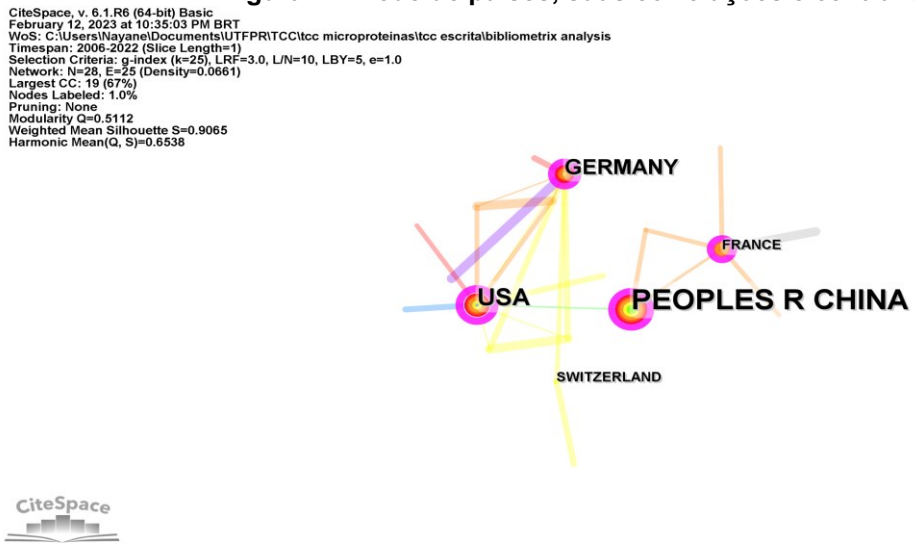
Figura 13: Artigos publicados pelos países ao longo dos anos.



Fonte: Criado com a ferramenta Bibliometrix do R versão 4.2.1 (ARIA; CUCCURULLO, 2017).

Os dados sobre os autores indicam baixa correlação, centralidade (Figura 9) e baixa colaboração entre eles (Figura 10). China, Estados Unidos, Alemanha, França e Suíça são os países que influenciam na área e para outros locais a centralidade foi 0 (Figura 14). Esses dados combinados aos artigos publicados ao longo dos anos (Figura 13) conversam entre si, sendo a China o país com grande destaque.

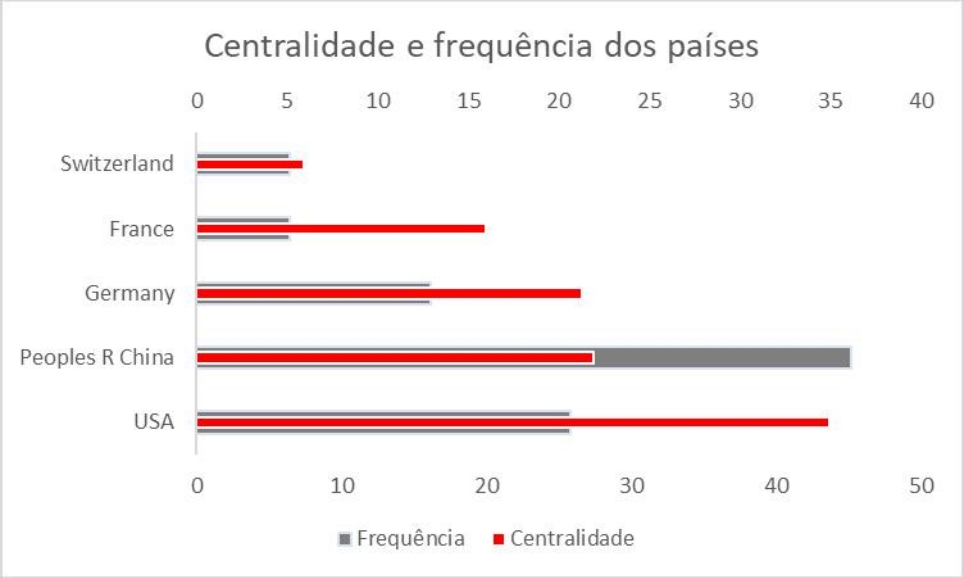
Figura 14: Rede de países, suas correlações e centralidade.



Fonte: Criado no CiteSpace (CHEN; DUBIN; KIM, 2014).

A influência indicada pela centralidade nem sempre está associada a um maior número de publicações, no gráfico 4 observamos um comportamento diferente da China, a qual possui um grande número de publicações (28) que corrobora com os dados das instituições, mas centralidade de apenas 0.37 que é muito próxima a da Alemanha (0.36) sendo que esta possui somente 10 publicações neste contexto. No entanto, o comportamento dos Estados Unidos é o oposto do observado para China, este país é quem mais influencia nesta área com centralidade de 0.59 e apenas 16 publicações 16, ou seja, 43% menos publicações e 37% menos centralidade comparado a China. Esse comportamento pode ser um indicativo de qualidade do trabalho, mas também temas de publicações de interesse local.

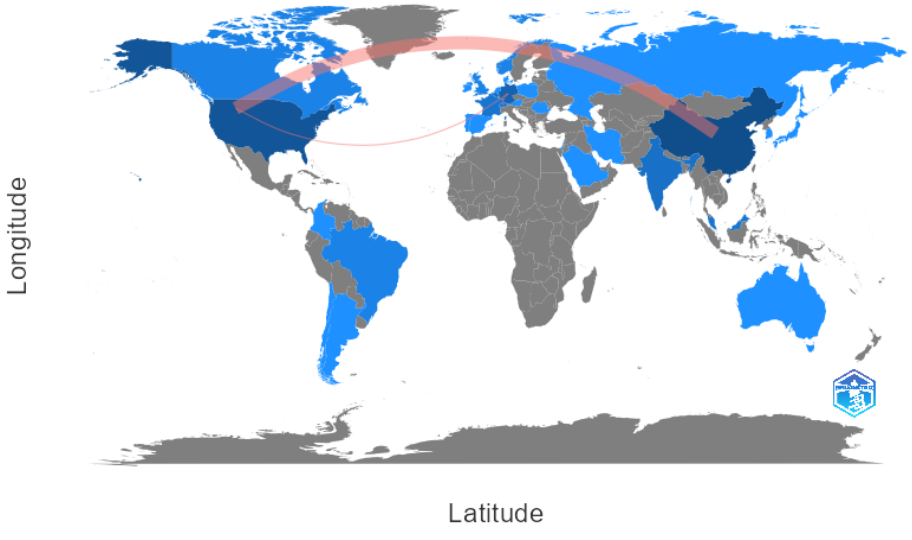
Gráfico 3: Gráfico representativo da centralidade e frequência de publicações de países.



Fonte: Dados obtidos no CiteSpace (CHEN; DUBIN; KIM, 2014)

O mapa de colaborações (figura 15) nos traz a imagem do que vimos com os outros dados, maior número de publicações da China e Estados Unidos, sendo que esses países são os que possuem maior influência e mais colaboram, neste mapa fica clara a centralização de publicações em poucos grupos de pesquisa.

Figura 15: Eixo de colaboração mundial de acordo com os 36 artigos sobre base de dados.



Fonte: Criado no Bibliometrix (ARIA; CUCCURULLO, 2017).

Os dados relacionados às instituições e países corroboram entre si, e indicam maior qualidade e maior interesse nos artigos dos Estados Unidos devido à relação inversamente proporcional entre centralidade e número de publicações. Essa área ainda é pouco explorada mundialmente e vimos que a partir de 2016 teve um aumento de trabalhos nesses poucos países (Figura 13), ou seja, o tema é de interesse da comunidade científica devido ao aumento, mas infelizmente restrito à poucas regiões.

5 CONSIDERAÇÕES FINAIS

As pequenas ORFs e proteínas traduzidas a partir delas representam um campo que ainda há muito a ser explorado e um espaço para expandir as contribuições científicas com outros países. O grande investimento em pesquisa é um fator importante quando falamos de produção acadêmica, técnicas de Ribo-Seq, RNA-seq e MS são relativamente caras e gerar os próprios dados para análise pode custar milhares de dólares. Felizmente, quando consideramos a bioinformática nos deparamos com um grande número de dados disponíveis para serem analisados e, apesar do investimento em bons computadores, os valores são menores que laboratórios, equipamentos e kits para realizar determinadas técnicas.

Com relação aos termos, encontramos três áreas de investigação interessantes, câncer e peptídeos antimicrobianos que aparecem frequentemente nos trabalhos lidos devido a sua importância na saúde da população e, curiosamente, foi identificado uma classe de peixes associada a esse tipo de pesquisa e levanta a questão sobre qual a importância desse grupo específico no contexto de sORFs.

O papel de colaborações é indubitavelmente importante, pois permite que novas perguntas, novas ideias e novas hipóteses sejam elaboradas. No entanto, na área de sORFs em bioinformática essas conexões são quase inexistentes tanto entre autores quanto países, poucos pesquisadores conversam entre si e somente cinco países aparecem com visibilidade na área, mesmo com o crescimento no número de publicações desde 2016.

Apesar das poucas colaborações e um número relativamente baixo de publicações, o índice H dos 82 estudos incluídos é 21, um número que assinala certa relevância deste tópico e que há muito espaço para ainda ser explorado.

Por fim, esse estudo cienciométrico sobre sORFs e pequenas proteínas revela uma área com potencial para crescimento e que está associada a tópicos emergentes de grande interesse mundial como câncer e tratamento de infecções multirresistentes. Embora o Brasil não apareça no pequeno grupo de países relevantes para área, a possibilidade de trabalhar com dados públicos sugere que grandes pesquisas possam ser feitas acessando esses repositórios e tratando esses dados computacionalmente.

REFERÊNCIAS BIBLIOGRÁFICAS

ADHIKARI, S. *et al.* A high-stringency blueprint of the human proteome. **Nature communications**, England, v. 11, n. 1, p. 5301, 2020.

ARIA, M.; CUCCURULLO, C. bibliometrix: An R-tool for comprehensive science mapping analysis. **Journal of Informetrics**, [*s. l.*], v. 11, n. 4, p. 959–975, 2017. Disponível em: https://www.sciencedirect.com/science/article/pii/S1751157717300500. Acesso em: 10 fev. 2023.

BASHIR, K. *et al.* Transcriptomic analysis of rice in response to iron deficiency and excess. **Rice (New York, N.Y.)**, United States, v. 7, n. 1, p. 18, 2014.

BLANVILLAIN, R. *et al.* The Arabidopsis peptide kiss of death is an inducer of programmed cell death. **The EMBO journal**, England, v. 30, n. 6, p. 1173–1183, 2011.

BLÜMEL, C.; SCHNIEDERMANN, A. Studying review articles in scientometrics and beyond: a research agenda. **Scientometrics**, [*s. l.*], v. 124, n. 1, p. 711–728, 2020. Disponível em: https://doi.org/10.1007/s11192-020-03431-7. Acesso em: 05 dez. 2022

BURRELL, Q. L. Hirsch's h-index: A stochastic model. **Journal of Informetrics**, [*s. l.*], v. 1, n. 1, p. 16–25, 2007. Disponível em: https://www.sciencedirect.com/science/article/pii/S1751157706000046. Acesso em: 05 dez. 2022.

CASSON, S. A. *et al.* The POLARIS gene of Arabidopsis encodes a predicted peptide required for correct root growth and leaf vascular patterning. **Plant Cell**, Integrative Cell Biology Laboratory, School of Biological and Biomedical Sciences, University of Durham, South Road, Durham DH1 3LE, United Kingdom., v. 14, n. 8, p. 1705–1721, 2002.

CHEN, C. M. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. **JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY**, Drexel Univ, Coll Informat Sci & Technol, Philadelphia, PA 19104 USA, v. 57, n. 3, p. 359–377, 2006.

CHEN, L. *et al.* The Small Open Reading Frame-Encoded Peptides: Advances in Methodologies and Functional Studies. **ChemBioChem**, State Key Laboratory of Chemical Biology and Drug Discovery, Department of Applied Biology and Chemical Technology, Hong Kong Polytechnic University, Hung Hom, 999077, Hong Kong Laboratory for Synthetic Chemistry and Chemical Biology Limited, Hong Kong Sc, v. 23, n. 8, 2022. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85120682139&doi=10.1002%2Fcbic.202100534&partnerID=40&md5=870374bf4153677619f5b99a925557dc. Acesso em: 11 dez. 2022.

CHEN, C.; DUBIN, R.; KIM, M. C. Emerging trends and new developments in regenerative medicine: a scientometric update (2000 - 2014). **Expert opinion on biological therapy**, England, v. 14, n. 9, p. 1295–1317, 2014.

CHEN, C.; SONG, M. Visualizing a field of research: A methodology of systematic scientometric reviews. **PLoS ONE**, [*s. l.*], v. 14, n. 10, 2019. Disponível em: http://dx.doi.org/10.1371/journal.pone.0223994. Acesso em 10 fev.2023

COMBIER, J.; LAURES-SERGUES, D.; BECARD, G. **Use of micropeptides for promoting plant growth**. Concessão: 2020.

COUSO, J.; PATRAQUIM, P. Classification and function of small open reading frames. **MOLECULAR CELL BIOLOGY**, [*s. l.*], v. 18, n. 9, p. 575–589, 2017.

DE CONINCK, B. *et al.* Mining the genome of Arabidopsis thaliana as a basis for the identification of novel bioactive peptides involved in oxidative stress tolerance. **J Exp Bot**, Centre for Microbial and Plant Genetics, KU Leuven, 3001 Heverlee, Belgium., v. 64, n. 17, p. 5297–5307, 2013.

DINGER, M. E. *et al.* Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities. **PLOS Computational Biology**, [*s. l.*], v. 4, n. 11, p. e1000176, 2008. Disponível em: https://doi.org/10.1371/journal.pcbi.1000176. Acesso em 09 fev. 2023.

DJAKOVIC, S. *et al.* BRICK1/HSPC300 functions with SCAR and the ARP2/3 complex to regulate epidermal cell shape in Arabidopsis. **Development (Cambridge, England)**, England, v. 133, n. 6, p. 1091–1100, 2006.

DONG, X. *et al.* Zm908p11, encoded by a short open reading frame (sORF) gene, functions in pollen tube growth as a profilin ligand in maize. **Journal of experimental botany**, England, v. 64, n. 8, p. 2359–2372, 2013.

DRAGOMIR, M. P. *et al.* FuncPEP: A Database of Functional Peptides Encoded by Non-Coding RNAs. **NON-CODING RNA**, Univ Texas MD Anderson Canc Ctr, Dept Translat Mol Pathol, Houston, TX 77030 USA Carol Davila Univ Med & Pharm, Fundeni Clin Hosp, Dept Surg, Bucharest 022328, Romania Univ Texas MD Anderson Canc Ctr, Dept Bioinformat & Computat Biol, Houston, TX 77030 US, v. 6, n. 4, 2020.

FABRE, B.; COMBIER, J. P.; PLAZA, S. Recent advances in mass spectrometry–based peptidomics workflows to identify short-open-reading-frame-encoded peptides and explore their functions. **Current Opinion in Chemical Biology**, Laboratoire de Recherche en Sciences Végétales, UMR5546, Université de Toulouse, UPS, CNRS, Auzeville-Tolosane, 31320, France, v. 60, p. 122–130, 2021. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85098651671&doi=10.1016%2Fj.cbpa.2020.12.002&partnerID=40&md5=17c7408e7464d78633a718ac6a80ca06. Acesso em 10 fev. 2023.

FIUME, E.; FLETCHER, J. C. Regulation of Arabidopsis embryo and endosperm development by the polypeptide signaling molecule CLE8. **The Plant cell**, England, v. 24, n. 3, p. 1000–1012, 2012.

FUKUDA, H. Signaling, transcriptional regulation, and asynchronous pattern formation governing plant xylem development. **Proceedings of the Japan Academy. Series B, Physical and biological sciences**, Japan, v. 92, n. 3, p. 98–107, 2016.

GUERRA-ALMEIDA, D.; NUNES-DA-FONSECA, R. Small Open Reading Frames: How Important Are They for Molecular Evolution?. **Frontiers in genetics**, Switzerland, v. 11, p. 574737, 2020.

GUERRA-ALMEIDA, D.; TSCHOEKE, D. A.; NUNES-DA-FONSECA, R. Understanding small ORF diversity through a comprehensive transcription feature classification. **DNA Research**, [*s. l.*], v. 28, n. 5, p. 1–18, 2021.

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

HARA, K. *et al.* The secretory peptide gene EPF1 enforces the stomatal one-cell-spacing rule. **Genes & development**, United States, v. 21, n. 14, p. 1720–1725, 2007.

HIGUCHI-TAKEUCHI, M. *et al.* Effect of small coding genes on the circadian rhythms under elevated CO(2) conditions in plants. **Plant molecular biology**, Netherlands, v. 104, n. 1–2, p. 55–65, 2020.

HUANG, J.-Z. *et al.* A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. **Molecular cell**, United States, v. 68, n. 1, p. 171-184.e6, 2017.

HUANG, H.-H. *et al.* Long non-coding RNA HOXB-AS3 promotes myeloid cell proliferation and its higher expression is an adverse prognostic marker in patients with acute myeloid leukemia and myelodysplastic syndrome. **BMC cancer**, England, v. 19, n. 1, p. 617, 2019.

ICHIHARA, K.; NAKAYAMA, K. I.; MATSUMOTO, A. Identification of unannotated coding sequences and their physiological functions. **The Journal of Biochemistry**, [*s. l.*], v. 00, n. August, p. 1–6, 2022.

INGOLIA, N. T. *et al.* Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. **Science (New York, N.Y.)**, United States, v. 324, n. 5924, p. 218–223, 2009.

INGOLIA, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. **Nature Reviews Genetics**, [*s. l.*], v. 15, n. 3, p. 205–213, 2014. Disponível em: https://doi.org/10.1038/nrg3645. Acesso em: 15 fev. 2023.

KUTE, P. M. *et al.* Small Open Reading Frames, How to Find Them and Determine Their Function. **Front Genet**, Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway. Sars International Centre for Marine Molecular Biology, University of Bergen, Bergen, Norway. Department of Molecular Epidemiology Of Vascular and Brain Disorders, v. 12, p. 796060, 2021.

LAURESSERGUES, D. *et al.* Primary transcripts of microRNAs encode regulatory peptides. **Nature**, England, v. 520, n. 7545, p. 90–93, 2015.

LEASE, K. A.; WALKER, J. C. The Arabidopsis unannotated secreted peptide database, a resource for plant peptidomics. **PLANT PHYSIOLOGY**, Univ Missouri, Div Biol Sci, Columbia, MO 65211 USA, v. 142, n. 3, p. 831–838, 2006.

LEONG, A. Z. X. *et al.* Short open reading frames (sORFs) and microproteins: an update on their identification and validation measures. **Journal of Biomedical Science**, UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia, Kuala Lumpur, 56000, Malaysia Division of Biomedical Science, School of Pharmacy, University of Nottingham Malaysia, Semenyih, Selangor, 43500, Malaysia, v. 29, n. 1, 2022. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85126554802&doi=10.1186%2Fs12929-022-00802-5&partnerID=40&md5=d05d92373e6e53a8f3e6de810052791f. Acesso em 01 fev. 2023.

LUO, X. *et al.* SPENCER: A comprehensive database for small peptides encoded by noncoding RNAs in cancer patients. **Nucleic Acids Research**, School of Life Sciences, Precision Medicine Institute, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, 510080, China State Key Laboratory of Oncology in South China, Cancer Center, Collaborative Innovation Center for Cancer Medicine, Sun, v. 50, n. D1, p. D1373–D1381, 2022. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85123387722&doi=10.1093%2Fnar%2Fgkab822&partnerID=40&md5=8447649d0fe7d2198a0acb9362511738. Acesso em: 28 jan. 2023.

MEMARIANI, H. *et al.* Melittin: from honeybees to superbugs. **Applied microbiology and biotechnology**, Germany, v. 103, n. 8, p. 3265–3276, 2019.

MUDGE, J. M. *et al.* A community-driven roadmap to advance research on translated open reading frames detected by Ribo-seq. **bioRxiv**, [*s. l.*], p. 2021.06.10.447896, 2021. Disponível em: http://biorxiv.org/content/early/2021/06/10/2021.06.10.447896.abstract. Acesso em: 05 fev. 2023.

NANJO, Y. *et al.* Transcriptional responses to flooding stress in roots including hypocotyl of soybean seedlings. **Plant molecular biology**, Netherlands, v. 77, n. 1–2, p. 129–144, 2011.

NARITA, N. N. *et al.* Overexpression of a novel small peptide ROTUNDIFOLIA4 decreases cell proliferation and alters leaf shape in Arabidopsis thaliana. **The Plant journal : for cell and molecular biology**, England, v. 38, n. 4, p. 699–713, 2004.

OHKUBO, Y. *et al.* Shoot-to-root mobile polypeptides involved in systemic regulation of nitrogen acquisition. **Nature plants**, England, v. 3, p. 17029, 2017.

ONG, Sheue Ni *et al.* Small open reading frames in plant research: from prediction to functional characterization. **3 Biotech**, [*s. l.*], v. 12, n. 3, p. 1–16, 2022. Disponível em: https://doi.org/10.1007/s13205-022-03147-w. Acesso em 28 jan. 2023.

ONG, S N *et al.* Small open reading frames in plant research: from prediction to functional characterization. **3 Biotech**, Centre for Research in Biotechnology for Agriculture (CEBAR), Universiti Malaya, 50603 Kuala Lumpur, Malaysia. GRID: grid.10347.31. ISNI: 0000 0001 2308 5949 Institute of Biological Sciences, Faculty of Science, Universiti Malaya, 50603 Kuala Lumpur, Mala, v. 12, n. 3, p. 76, 2022.

ORR, M. W. *et al.* Alternative ORFs and small ORFs: shedding light on the dark proteome. **Nucleic acids research**, England, v. 48, n. 3, p. 1029–1042, 2020.

ROHRIG, H. *et al.* Soybean ENOD40 encodes two peptides that bind to sucrose synthase. **Proceedings of the National Academy of Sciences of the United States of America**, United States, v. 99, n. 4, p. 1915–1920, 2002.

SALISBURY, J. P. *et al.* The central nervous system transcriptome of the weakly electric brown ghost knifefish (Apteronotus leptorhynchus): de novo assembly, annotation, and proteomics validation. **BMC Genomics**, Barnett Institute, Department of Chemistry and Chemical Biology, Northeastern University, 360 Huntington Avenue, 412 TF, Boston, MA, 02115, USA. j.salisbury@neu.edu. Laboratory of Neurobiology, Department of Biology, Northeastern University, 360 Huntingto, v. 16, n. 1, p. 166, 2015.

STEINBERG, R.; KOCH, H. G. The largely unexplored biology of small proteins in pro- and eukaryotes. **FEBS Journal**, [*s. l.*], v. 288, n. 24, p. 7002–7024, 2021.

TABATA, R. *et al.* Perception of root-derived peptides by shoot LRR-RKs mediates systemic N-demand signaling. **Science (New York, N.Y.)**, United States, v. 346, n. 6207, p. 343–346, 2014.

WANG, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. **Nature**, [*s. l.*], v. 604, n. 7906, p. 437–446, 2022. Disponível em: https://doi.org/10.1038/s41586-022-04601-8. Acesso em: 29 jan. 2023.