UNIVERSIDADE FEDERAL DO MARANHÃO
Fundação Instituída nos termos da Lei 5.152 de 21/10/1966 – São Luís – Maranhão
CENTRO DE CIÊNCIAS BIOLÓGICAS E DA SAÚDE
COORDENADORIA DO CURSO DE CIÊNCIAS BIOLÓGICAS (Modalidade: licenciatura)

THAYANNE CHRISTTINE COSTA ARAÚJO

CARACTERIZAÇÃO DE PROTEÍNAS HIPOTÉTICAS DO GENOMA DA CIANOBACTÉRIA *PANTANALINEMA* sp. GBBB05 UTILIZANDO UMA ABORDAGEM COMPUTACIONAL

SÃO LUÍS

2022

THAYANNE CHRISTTINE COSTA ARAÚJO

CARACTERIZAÇÃO DE PROTEÍNAS HIPOTÉTICAS DO GENOMA DA CIANOBACTÉRIA *PANTANALINEMA* sp. GBBB05 UTILIZANDO UMA ABORDAGEM COMPUTACIONAL

Monografia apresentada ao curso de Ciências Biológicas da Universidade Federal do Maranhão, para obtenção do título de Licenciatura em Ciências Biológicas.

Orientador: Leonardo Teixeira Dall'Agnol.

SÃO LUÍS

2022

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Araújo, Thayanne Christtine Costa.
Caracterização de proteínas hipotéticas do genoma da cianobactéria Pantanalinema sp. GBBB05 utilizando uma abordagem computacional / Thayanne Christtine Costa Araújo. - 2022.
60 f.

Orientador(a): Leonardo Teixeira Dall'Agnol.
Monografia (Graduação) - Curso de Ciências Biológicas, Universidade Federal do Maranhão, São Luís, 2022.

1. Agrupamentos gênicos. 2. Bioinfomática. 3. Genoma. 4. Proteínas hipotéticas. I. Teixeira Dall'Agnol, Leonardo. II. Título.

THAYANNE CHRISTTINE COSTA ARAÚJO

CARACTERIZAÇÃO DE PROTEÍNAS HIPOTÉTICAS DO GENOMA DA CIANOBACTÉRIA *PANTANALINEMA* sp. GBBB05 UTILIZANDO UMA ABORDAGEM COMPUTACIONAL

Aprovada em: _____/_____/_____

BANCA EXAMINADORA

_____

Prof. Dr. Leonardo Teixeira Dall'Agnol – Orientador
Universidade Federal do Maranhão – UFMA

_____

Prof. Dr. Silvio Gomes Monteiro – 1° Examinador
Universidade Federal do Maranhão – UFMA

_____

Dr. Alex Ranieri Lima Gerônimo – 2° Examinador
Instituto Butantan

AGRADECIMENTOS

Não teria como não agradecer primeiro à minha família. À minha irmã **Raquel**, que é um exemplo de determinação, uma vez que trabalha e estuda, uma rotina que vemos o quanto a cansa, mas ela sempre tenta dar seu melhor e conciliar ambas as coisas. A meu pai **Aldenir** e minha mãe **Maria,** que sempre foram pais extremamente presentes e amorosos que sempre nos apoiaram e fizeram de tudo para que tivéssemos boa educação. Amo-lhes muito. Então, agradeço também a **Deus** por esta realização e pelas pessoas que colocou em minha vida, em especial pela minha família.

Aos meus amigos desde o tempo do fundamental **Iago**, **Juliana**, **João Victor** e **José Henrique** que sempre tiveram presentes, mesmo nos momentos mais difíceis, me apoiando e me proporcionando excelentes momentos. São pessoas muito especiais a quem admiro e amo muito.

Agradeço também aos meus amigos de jogatina que se tornarão mais que amigos de jogatina **Neemias**, **Daniel** e **Débora**. Se tornaram diariamente presentes na minha vida, sempre me oferecendo momentos de descontração e risadas. Pessoas que sei que posso sempre contar, que confio, que me ajudam sempre que preciso (mesmo sem saberem) e que amo.

E obviamente, não poderia esquecer os amigos que fiz durante a graduação. Então começo agradecendo ao **grupo GB3**. Aos professores **Leonardo** e **Hivana** que me deram a oportunidade de fazer parte do grupo, só tenho a agradecê-los por tudo que aprendi e às orientações que deram a mim e aos demais colegas. À **Anna Letícia** e ao **Isaías** que sempre estavam disponíveis para me tirarem dúvidas sobre bioinformática. À **Manu**, que foi uma excelente dupla de laboratório. E especialmente ao **Lucas Salomão**, que foi quem me aconselhou em um momento difícil na minha vida, eu estava completamente perdida no curso sem saber por onde seguir e ele me fez uma super propaganda do GB3 e me incentivou a tentar fazer parte do grupo, era o conselho que eu precisava naquele momento pra continuar a seguir no curso.

À **LAGeM** pelos anos de aprendizado que me ofereceu, eventos incríveis e bons companheiros. Foi muito bom fazer parte da liga!

À **Amanda Letícia** que sempre foi um exemplo de determinação e espontaneidade pra mim e que puxou minha orelha algumas vezes durante o período de estágio curricular. Ao

Leandro Robson que sempre contagiava a turma com sua alegria e brincadeiras e que também sempre foi exemplo, sempre dedicado e solidário.

Ao meu fiel amigo **Fabio Ferreira**, que esteve ao meu lado em vários momentos na graduação e na vida. Uma pessoa a quem admiro e gosto muito, sempre muito determinado e dedicado, me proporcionou várias conversas maravilhosas e divertidas e quem eu espero que sempre esteja presente em minha vida.

À **Mara Bogéa** e **Milena Alves** que sempre foram muito gentis comigo, me dando conselhos e abraços. São duas mulheres maravilhosas e exemplos pra mim, sempre dedicadas a aprender e a ouvir.

À **Alícia Brendha**, **Bruna Fernandes**, **Ingrid Monteiro** e **Gaby Ramos** por sempre terem me dado muito apoio, carinho, e terem sido amigas maravilhosas durante a graduação.

À **Bruna Martins**, sou muito grata por ter sido minha amiga e compartilhado comigo vários momentos de desespero, mas também momentos muito divertidos. Passamos por muitas disciplinas juntas e ela foi essencial em várias etapas da graduação, ajudávamos uma à outra, fazíamos vários trabalhos juntas e quando estávamos muito desesperadas, tínhamos crises de risos. Não consigo lembrar de nenhum momento feliz durante a graduação que a Bruna não estivesse presente.

Por fim, ao meu amor, **Victor Moreira**, que sempre acreditou em mim, sempre esteve ao meu lado e foi meu parceiro. O admiro muito pelo coração generoso que tem, pela pessoa boa, inteligente e prestativa que é. Obrigada por todas as vezes que me encorajou, que me ouviu ou que simplesmente ficou ao meu lado e por ser minha inspiração. Todas as vezes que acreditou em mim, quando eu mesma não acreditava. Foi o apoio e incentivo dele que me deram forças pra seguir na graduação e fazer este trabalho.

RESUMO

As cianobactérias, antigamente conhecidas como algas verdes azuis, são um filo procariótico muito antigo, que possui uma ampla diversidade morfológica, grande importância ecológica por fazerem parte do ciclo do oxigênio e nitrogênio, e produzirem diversos metabólitos secundários, os quais podem apresentar potencial terapêutico e biotecnológico. Os ambientes de água doce apresentam características mais favoráveis para o crescimento de cianobactérias, porém, existem espécies com capacidade de crescer em outros ambientes e adaptadas a uma ampla variação de pH. *Pantanalinema* é um gênero recente de cianobactérias pertencente da família *Leptolynbyaceae*, que vivem em águas salino-alcalinas do Pantanal, bioma brasileiro, a um pH cuja a taxa varia de 4 a 11. Recentemente, uma espécie do gênero foi descrita no cerrado, isolada de uma cachoeira localizada na orla externa da Chapada das Mesas, situada no município Carolina-MA, e teve seu genoma sequenciado pelo grupo de pesquisa em biodiversidade, bioprospecção e biotecnologia (GB3) da UFMA, sendo identificada como *Pantanalinema* sp. GBBB05. A bioinformática é uma ferramenta importante para interpretação, armazenamento e anotações de dados que são gerados a partir do sequenciamento genético. Através de ferramentas da bioinformática, é possível extrair informações importantes sobre um organismo utilizando as sequências de aminoácido ou nucleotídeos de um genoma. O trabalho tem como principal objetivo fazer uma caracterização de proteínas hipotéticas do genoma da *Pantanalinema* sp. GBBB05. A submissão do genoma no *antiSMASH*, para prospecção e mineração do genoma, identificou 80 proteínas hipotéticas distribuídas em 14 agrupamentos gênicos, as quais foram caracterizadas manualmente buscando identificar domínios e superfamílias, parâmetros físico-químicos, localização subcelular, atividade secretora e número de hélices transmembranares, utilizando ferramentas de bioinformática e bancos de dados como o Uniprot. Além disso, foram construídas árvores gênicas através do MEGAx, e análises de sintenia, com o *clinker* e *clustermap.js*, para entender a relação evolutiva com outros organismos. Através dos resultados obtidos, é possível dar passos iniciais para a caracterização genômica da *Pantanalinema* sp. GBBB05, demonstrando um possível potencial biotecnológico e contribuindo para a caracterização da diversidade local.

Palavras-chave: Genoma, bioinformática, proteínas hipotéticas, agrupamentos gênicos.

ABSTRACT

Cyanobacteria, formerly blue-green algae, are a very old prokaryotic phylum, which has a wide morphological diversity, great ecological importance because they are part of the oxygen and nitrogen cycle, and produce several secondary metabolites, which may have therapeutic and biotechnological potential. Freshwater environments present more favorable characteristics for the growth of cyanobacteria, however, there are species with the ability to grow in other environments and adapted to a wide range of pH. *Pantanalinema* is a recent genus of cyanobacteria belonging to the *Leptolynbyaceae* family, which live in saline-alkaline waters of the Pantanal, Brazilian biome, at a pH whose rate varies from 4 to 11. Recently, a species of the genus was described in the cerrado, isolated from a waterfall located on the outer edge of Chapada das Mesas, located in the municipality of Carolina-MA, and had its genome sequenced by the research group on biodiversity, bioprospecting and biotechnology (GB3) at UFMA, being identified as *Pantanalinema* sp. GBBB05. Bioinformatics is an important tool for interpreting, storing and annotating data generated from genetic sequencing. Through bioinformatics tools, it is possible to extract important information about an organism using the amino acid or nucleotide sequences of a genome. The main objective of this work is to characterize hypothetical proteins in the genome of *Pantanalinema* sp. GBBB05. Genome submission to *antiSMASH* for genome prospecting and mining identified 80 hypothetical proteins distributed in 14 gene clusters, which were manually characterized seeking to identify domains and superfamilies, physicochemical parameters, subcellular location, secretory activity and the number of transmembrane helices, using bioinformatics tools and databases such as Uniprot. Furthermore, gene trees were constructed using MEGAx, and synteny analysis with clinker and clustermap.js, to understand the evolutionary relationship with other organisms. Through the results obtained, it is possible to take initial steps for the genomic characterization of Pantanalinema sp. GBBB05, demonstrating a possible biotechnological potential and contributing to the characterization of local diversity.

Keywords: Genome, bioinformatics, hypothetical proteins, gene clusters.

SUMÁRIO

1.INTRODUÇÃO

O Parque Nacional da Chapada das Mesas – Maranhão (PNCM), possui grande importância ecológica, hidrológica e biotecnológica, além de ter um grande potencial turístico. Apesar disso, existem poucos estudos sobre a comunidade de cianobactérias que o habitam. Assim, pesquisas que objetivem caracterizar essas comunidades, realizando estudos com mineração genômica de vias metabólicas de interesse biotecnológicos são importantes e necessários, contribuindo para a compreensão da diversidade microbiana desse ecossistema e elucidando um potencial de aspectos biotecnológicos ainda desconhecido.

O genoma da cianobactéria *Pantanalinema* sp. GBBB05, que foi isolada de uma cachoeira localizada na orla externa da Chapada das Mesas, é pouco anotado automaticamente, possuindo muitos genes e proteínas não caracterizados, demonstrando uma lacuna na anotação automática. O trabalho foi focado na mineração do genoma e prospecção de vias metabólicas que possam ter algum potencial biotecnológico, objetivando caracterizar funcionalmente proteínas hipotéticas presentes nessas vias metabólicas por meio de ferramentas e servidores de bioinformática.

O avanço no sequenciamento genético durante os últimos anos gerou uma grande quantidade de dados. As sequências dos genomas são traduzidas em sequências de proteínas e muitas dessas sequências já são conhecidas estrutural e funcionalmente, encontrando-se armazenadas em bancos de dados que foram desenvolvidos com esta finalidade. (THAKUR et al., 2020).

Em resposta ao grande número de pesquisas envolvendo sequenciamento de genomas inteiros, surgem várias sequências de proteínas e DNA cujas funções são desconhecidas. Essas sequências são anotadas em bancos de dados, recebendo o nome de “proteínas hipotéticas” (HPs). As ferramentas bioinformáticas auxiliam na caracterização dessas HPs, possibilitando que haja uma maior compreensão dos genes e das proteínas de um determinado genoma (SILVA, F. F.; GONÇALVES, D. B.; LOPES, 2020).

A bioinformática é uma ferramenta importante para interpretação, armazenamento e anotações de dados que são gerados a partir do sequenciamento genético, ajudando, também, a atribuir um sentido às sequências de DNA ou proteínas. (PEVSNER, 2015). Por meio destas informações, pode-se identificar alvos para compreender patologias, reconhecer compostos de valor industrial, desenvolver fármacos e disponibilizar estas informações em vários bancos de dados (SOUZA, L. N.; RHODEN, S. A.; PAMPHILE, 2014).

2.REFERENCIAL TEÓRICO

2.1 As cianobactérias e a mineração genômica

As cianobactérias, ou algas azuis, são um filo procariótico muito antigo e que possui uma ampla diversidade morfológica. Atribui-se a esses organismos, através da realização da fotossíntese, o início do desenvolvimento de uma atmosfera composta por oxigênio há bilhões de anos atrás (SCHIRRMEISTER; ANTONELLI; BAGHERI, 2011). Elas têm ampla distribuição e a diversidade desse filo deve-se à adaptação ecológica que possuem, a qual é facilitada por vários fatores dentre os quais podemos citar a reprodução assexuada e especiação devido a pressões ecológicas (WILLIS; WOODHOUSE, 2020)

Além disso, as cianobactérias possuem grande importância ecológica por fazerem parte dos ciclos biogeoquímicos do carbono, nitrogênio e do oxigênio. Estes microrganismos possuem também a capacidade de produzir diversos metabólitos secundários, dentre eles estão as toxinas. As cianotoxinas abrangem diversas classes químicas e de efeitos variados como as microcistinas (hepatotoxinas), as anatotoxinas e saxitoxinas (neurotoxinas), por exemplo (DITTMANN, E.; NEILAN, B.; BÖRNER, 2001; SANCHES et al., 2012). Devido ao fato dos metabólicos secundários produzidos pelas cianobactérias serem altamente tóxicos, episódios de intoxicação de humanos e animais devido ao contato com essas toxinas são comuns em várias regiões. Isso se tornou uma questão de saúde pública, chamando a atenção dos órgãos de gestão das águas e saúde pública, uma vez que, sob condições favoráveis em ambientes de água doce, algumas cianobactérias conseguem formar florações, crescimento populacional desses microrganismos, que interfere diretamente na qualidade da água (JESUS et al., 2010), o que foi observado frequentemente no reservatório Joanes I da região metropolitana de Salvador, Bahia (MENESCAL, 2018).

Como existe uma grande diversidade de metabólitos secundários produzidos pelas cianobactérias, alguns podem ter potencial terapêutico e biotecnológico. Algumas substâncias toxinas ou produtos bioativos produzidos por esses microrganismos podem ser utilizados como

aleloquímicos, demonstrando potencial para desenvolvimento e aplicações desses compostos como algicidas, inseticidas e herbicidas (BERRY, 2008). Além de também já terem sido descritas toxinas com papel importante no tratamento de doenças, até mesmo alguns tipos de câncer, por possuírem potencial antioxidante (GUERREIRO, 2019), possuindo compostos derivados do seu metabolismo que exibem ações antibacterianas, antitumoral, imunossupressoras, anti-HIV e antifúngica (VIJAYAKUMAR; MENAKHA, 2015).

Por serem fotossintetizantes, as cianobactérias conseguem fixar CO_2 a partir da energia solar. Isso as torna uma alternativa mais vantajosa, econômica e ecológica para produção de biocombustível e produtos químicos a partir da conversão biológica direta do CO_2 nessas substâncias, tendo alto potencial para substituir os combustíveis fósseis (MACHADO, I. M. P.; ATSUMI, 2012; MILANO et al., 2016)

Os ambientes de água doce apresentam características mais favoráveis para o crescimento de cianobactérias, pois a maioria das espécies demonstram um melhor crescimento em águas com pH 6 a 9, temperatura entre 15° a 30° e que possuem elevada concentração de nutrientes (FUNASA, 2003). Porém, existem espécies com capacidade de crescer em outros ambientes e adaptadas a uma ampla variação de pH. *Pantanalinema* é um gênero recente de cianobactérias pertencente da família *Leptolynbyaceae*, que vivem em águas salino-alcalinas do Pantanal, bioma brasileiro, a um pH cuja a taxa varia de 4 a 11. Além de conseguirem viver nesse meio, quando colocadas em um meio de cultura, produziram biomassa e foram capazes de alterar o pH do meio (VAZ et al., 2015). Recentemente ela foi descrita no ambiente do Cerrado e teve seu genoma sequenciado pelo grupo de pesquisa em biodiversidade, bioprospecção e biotecnologia (GB3) da UFMA demonstrando seu potencial biotecnológico (FERREIRA et al., 2021).

2.2 Bancos de dados

Com o grande volume de informação gerado nos últimos anos, os bancos de dados são importantes para armazenar várias informações sobre dados biológicos, mantendo acessíveis e funcionais as sequências de anotações genômicas, informações sobre proteínas, metabólitos,

entre outros. Os bancos de dados biológicos se tornaram, assim, um importante meio para estudos na área de bioinformática, facilitando consultas e a atualização dos dados (LIBÓRIO; RESENDE, 2021).

Um dos bancos de dados mais utilizados atualmente é o *Genbank*, construído e mantido pelo NCBI (*National Center for Biotechnology Information*), que incorpora sequências de DNA de milhares de organismos diferentes, com todos os dados de acesso público (BENSON, 2000). Outro banco de dados bastante útil para a comunidade científica é o Uniprot (BATEMAN et al., 2015), que tem como objetivo fornecer aos usuários um amplo conjunto de sequências de proteínas de alta qualidade anotadas com informações funcionais. Estima-se que o número de sequências disponibilizadas no Uniprot seja de aproximadamente 190 milhões (BATEMAN et al., 2021).

2.3 Caracterização automática e manual de proteínas hipotéticas

O genoma de um microrganismo consiste em centenas de milhares de regiões denomidas fase abertas de leituras (ORFs, *open reading frames*), que são sequências de DNA ou RNA as quais podem ser traduzidas em um polipeptídeo. Um computador é capaz de identificar essas ORFs e atribuir possíveis funções a elas, porém, nem todas essas regiões são caracterizadas, e muitas vezes codificam proteínas que não possuem homologia significativa com sequências de aminoácidos já conhecidas, sendo caracterizadas automaticamente em banco de dados como proteínas hipotéticas (MADIGAN, 2016).

A caracterização ou anotação manual de uma proteína hipotética tem como objetivo atribuir funções a essa proteína. O pesquisador irá utilizar seus conhecimentos e resultados gerados pela anotação automática para atribuir função a essas sequências, utilizando, principalmente, vários bancos de dados e ferramentas que irão executar diferentes análises para buscar definir uma funcionalidade à proteína hipotética (SOUZA, 2014).

2.4 Prospreccção de vias metabólicas pelo antiSMASH

Vários metabólicos secundários como terpenos, alcalóides e peptídeos, são substâncias sintetizadas por muitas enzimas multifuncionais, a exemplo do peptídeo síntese não ribossômico (NRPS), bem como as sintases de policetídeos tipo I e tipo II (PKS), ainda sendo

possível uma junção dos dois sistemas, além disso, peptídeos sintetizados por ribossomos e modificados pós-tradução (RiPPs) (WASE; WRIGHT, 2008).

O servidor do *antiSMASH* (*antibiotics and secondary metabolite analysis shell*), realiza uma abordagem baseada em regras para fazer uma mineração dos genomas submetidos, buscando diferentes tipos de vias relacionadas à biossíntese de metabólicos secundários e substâncias bioativas. As "regras" utilizadas pelo *antiSMASH* definem quais funções biossintéticas centrais devem existir em uma determinada região genômica para caracterizá-la como um agrupamento de genes biossintéticos de metabólitos secundários especializados. (BLIN et al., 2021).

2.5 Filogenia Molecular e filogenômica

O estudo da filogenia é importante para entender diversas questões biológicas, como as relações entre espécies ou genes. O constante avanço das técnicas de sequenciamento genético contribuiu para a evolução de análises filogenéticas, uma vez que surgiu uma infinidade de métodos filogenéticos e softwares que podem ser usados para a construção da análise (YANG; RENNALA, 2012). A comparação de dados de sequências moleculares é importante para, além de reconstruir as histórias evolutivas das espécies, inferir a natureza e a extensão das forças seletivas que atuam na evolução de genes e consequentemente na evolução de espécies (TAMURA et al., 2011).

Dentre os vários métodos utilizados para construção de árvores, têm-se o método *Neighbor-Joining* (SAITOU; NEI, 1987), com análises computacionais rápidas e de alta precisão ideais para a construção de árvores baseadas em dados de sequências de DNA ou proteínas (KUMAR; GADAGKAR, 2000). Além disso, existem os métodos de reamostragem, como o *bootstrap*, que gera novas amostras a partir de um conjunto de dados original com o objetivo de construir intervalos de confiança e testar uma hipótese (LIMA, 2017; SOUSA, 2018).

2.5 A análise de sintenia

A sintenia é um tipo de análise de genômica comparativa que tem como objetivo entender como dados genômicos estão relacionados em diferentes espécie**s** (CARNEIRO; COIMBRA, 2010)**.** A partir desse estudo, é possível entender como os genes estão organizados

em um genoma, ou agrupamento gênico, e observar as modificações que podem ter sofrido ao longo do tempo (PROSDÓCIMI; MOREIRA, 2015), podendo revelar a manutenção desses agrupamentos e de sequências regulatórias que são importantes em determinadas regiões genômicas (STRAALEN; ROELOFS, 2006)

3.OBJETIVOS

3.1 Objetivo Geral

Anotação e análise in silico das proteínas hipotéticas do genoma da *Pantanalinema* sp. GBBB05.

3.2 Objetivos Específicos

- Caracterizar as proteínas hipotéticas;
- Realizar a curadoria manual do genoma para otimização da anotação automática;
- Construir árvores gênicas utilizando sequência de aminoácidos de genes biossintéticos centrais, verificando a similaridade e a homologia dessas sequências com a de outras cianobactérias;
- Comparar alguns agrupamentos gênicos de interesse através da análise de sintenia.

4.METODOLOGIA

O genoma da linhagem *Pantanalinema* sp. GBBB05 foi obtido do *GenBank* a partir do número de acesso GCA_016743235.1 (FERREIRA et al., 2021).

As sequências das proteínas hipotéticas que foram obtidas nos agrupamentos de genes e vias metabólicas aferidas no *antiSMASH* versão 6.0 (BLIN et al., 2021) foram submetidas em diferentes bancos de dados e servidores online para realizar sua caracterização.

Além disso, foram construídas ávores gênicas através das sequências de proteínas codificadas por genes biossintéticos centrais e geradas análises de sintenia de alguns agrupamentos aferidos no *antiSMASH.*

4.1 Caracterização de proteínas hipotéticas (HPs)

4.1.1 Predição de domínios ou famílias de proteínas para descobrir a função

Para identificar essas regiões, foram utilizados bancos de dados como o BLASTp (MCGINNIS; MADDEN, 2004), CDD (Conserved Domain Database) (MARCHLER-BAUER et al., 2015), Pfam (MISTRY et al., 2021), SMART (*Simple Modular Architecture Research Tool*) (LETUNIC; 14 DOERKS; BORK, 2015; CATH (*Class Architecture Topology Homology*) (SILLITOE et al., 2015); Uniprot e InterPRO (MITCHELL et al., 2019).

Para visualização das proteínas hipotéticas no genoma da *Pantanalinema* sp. GBBB05 foi construído um mapa genômico através do CG*View* (STOTHARD; WISHART, 2005).

4.1.2 Análise de propriedades físico-químicas

O servidor *Expasy's ProtParam* foi utilizado para realização de análises de parâmetros físico-químicos das proteínas hipotéticas (GASTEIGER et al., 2005). A medida de dados teóricos de vários parâmetros físico-químicos, como massa molecular, coeficiente de extinção, ponto isoelétrico, índice de instabilidade, índice alifático e a grande média de hidropaticidade (GRAVY) foram evidenciados pela predição feita pelo servidor online.

4.1.3 Localização sub-celular

A fim de predizer a função de uma proteína em nível celular, precisa-se estimar a sua localização sub-celular (citoplasma, periplasma, membrana interna, membrana externa ou extracelular).

Para caracterizar os grupamentos gênicos e auxiliar na identificação de proteínas foram utilizados: PSORTb versão 3.0.3 (YU et al., 2010), PSLpred (BHASIN, GARG, RAGHAVA, 2005) e CELLO (YU et al., 2006), utilizados para predizer a localização sub-celular; o *SecretomeP 2.0 Server* (BENDTSEN et al., 2005), foi usado para prever a secreção de proteína não clássica, isto é, secreção de peptídeo sinal independente e a ferramenta SignalP 6.0 Server (ALMAGRO ARMENTEROS et al., 2019) usada para prever peptídeos sinais.

4.1.4 Hélice transmembrana e previsão de topologia

Os servidores *Deep*TMHMM (*Hidden Markov Model for Transmembrane Helices*) (HALLGREN et al., 2022) e CCTOP (*consensus restricted topology prediction server*)

(DOBSON; REMÉNYI; TUSNÁDY, 2015), foram utilizados para prever o número de regiões transmembranares; o servidor SOSUI (HIROKAWA; BOON-CHIENG; MITAKU, 1998) foi usado para fazer a classificação das proteínas em solúveis ou proteínas de membrana.

4.2 Avaliação de desempenho

Para avaliar a confiabilidade da previsão, foi utilizada a curva ROC (curva de característica de operação do receptor), através do *Online ROC CURVE Calculator*. Para realizar a análise, foi empregada a escala de classificação ordinal, a qual cada linha representa um caso sendo composta por dois números. O primeiro é 0 ou 1, onde 0 corresponde a um verdadeiro negativo e 1 a um verdadeiro positivo. O segundo número indica a classificação de confiança para cada caso, variando de 1 a 6 (ENG, 2014).

4.3 Filogenia Molecular e filogenômica

A árvore para análise de filogenia foi inferida pelo FastME 2.1.6.1 (LEFORT; DESPER; GASCUEL, 2015) baseada no proteoma inteiro, utilizando o TYGS (*type strain genome server*) (MEIER-KOLTHOFF; GÖKER, 2019)**.** Os genomas utilizados para a construção da árvore foram baixados do NCBI em formato FASTA.

Para construir as árvores gênicas, para filogenética molecular, foram selecionadas proteínas biossintéticas de alguns agrupamentos específicos. As sequências de aminoácidos das proteínas foram submetidas no BLASTp, onde foram selecionadas outras sequências de proteínas presentes no genoma de outras cianobactérias. As sequencias de aminoácidos das proteínas foram baixadas em arquivo FASTA e alinhadas utilizando o *Muscle* no MEGAx versão 11.0 (TAMURA; STECHER; KUMAR, 2021). Após o alinhamento, ainda no MEGAx, foram geradas as árvores pelo método de *Neighbor-Joinning* empregando o teste *bootstrap* com 1000 replicatas para avaliar a confiabilidade de cada nó.

4.4 Análise de Sintenia

Para realizar a análise, os agrupamentos de interesse foram baixados em formato *genbank* (gbk) do *antiSMASH*. Foram utilizados os programas *clinker* para realizar o alinhamento entres os agrupamentos e *clustermap.js* para gerar a imagem de visualização dos

resultados (GILCHRIST; CHOOI, 2021). As primeiras análises foram construídas com clusters conhecidos indicados pelo *antiSMASH*, em seguida alguns agrupamentos foram comparados com agrupamentos de espécies que apresentaram homologia com a espécie da *Pantanalinema* nas árvores gênicas construídas. Esses agrupamentos específicos tiveram seu código de acesso pesquisados no *genbank* e, após submetidos no *antiSMASH*, também foram baixados em formato gbk para construção da análise.

5. RESULTADOS E DISCUSSÃO

O genoma da *Pantanalinema* sp. GBBB05 tem 94 *contigs* e cerca de 7,181,771 bp (FERREIRA et al., 2021). Ao ser submetido no RAST (*Rapid Annotations using Subsystems Technology*) (AZIZ et al., 2008), foram identificadas 3530 são proteínas hipotéticas, o que corresponde a aproximadamente 50% do seu genoma. Porém, este trabalho focou na mineração e prospecção de vias metabólicas do genoma que possam ter um potencial biotecnológico.

5.1 Caracterização de proteínas hipotéticas

Através da submissão do genoma da *Pantanalinema* sp. GBBB05 no *antiSMASH* versão 6.0, foi possível obter 14 agrupamentos gênicos, dentre os quais cinco agrupamentos eram terpenos; três agrupamentos NRPS; um agrupamento betalactona; um agrupamento resorcinol; um agrupamento NRPS-T1PKS; um agrupamento NRPS, NRPS-*like*; um agrupamento contendo RRE (elemento de reconhecimento RiPP) e um agrupamento lantipeptídeo-classe-V.

Selecione uma região genômica:

Visão geral | 1.1 | 2.1 | 3.1 | 3.2 | 5.1 | 8.1 | 10.1 | 10.2 | 18.1 | 21.1 | 40,1 | 49,1 | 50,1 | 62.1

Identificou regiões de metabólitos secundários usando rigor 'relaxado'

Região	Modelo	A partir de	Para	Cluster conhecido mais semelhante		Semelhança
Região 1.1	terpeno	1	15.713			
Região 2.1	betalactona	221.361	245.891			
Região 3.1	resorcinol	1	30.652			
Região 3.2	terpeno	197.847	218.851			
Região 5.1	NRPS , T1PKS	204.834	239.621	aranazol A / aranazol B / aranazol C / aranazol D	NRP + policetídeo	12%
Região 8.1	NRPS	15.748	62.374			
Região 10.1	NRPS	1	52.895	nostopeptólido A2	Polipeptídeo + NRP: depsipeptídeo cíclico	37%
Região 10.2	terpeno	81.537	102.469			
Região 18.1	terpeno	36.163	58.082			
Região 21.1	NRPS-like , NRPS	78	108.460	puwainaficina F / minutissamida A	NRP	33%
Região 40.1	contendo RRE	37.870	58.118			
Região 49.1	terpeno	11.458	32.687			
Região 50.1	lanthipeptídeo-class-v	16.162	41.302			
Região 62.1	NRPS	1	28.438	nostociclopeptídeo A2	NRP	28%

Figura 1. Agrupamentos gênicos recuperados via *antiSMASH* para análise de genes de biossíntese de metabólitos secundários.

Ao todo, foram obtidas 80 proteínas hipotéticas, sendo que a maior quantidade de HPs se concentram nos agrupamentos terpeno e NRPS, conforme demonstra a tabela 1.

Tabela 1. Distribuição de HPs por agrupamento gênico.

Terpeno	NRPS	Betalactona	Resorcinol	NRPS, T1PKS	NRPS, NRPS-like	Contendo RRE	Lantipeptídeo - classe V	TOTAL
21	18	6	7	10	11	1	6	**80**

5.1.1 Predição de domínios e famílias de proteínas

As análises iniciais dessas proteínas, feitas com o auxílio dos preditores Pfam, CATCH, SMART, Interpro e CDD, tinha como principal objetivo identificar possíveis domínios ou famílias. Os domínios proteicos são módulos estruturais distintos, com funções características. A identificação de domínios e famílias com funções conhecidas pode ajudar na compreensão da função de uma determinada proteína, principalmente quando esta é uma proteína ainda não caracterizada (MADIGAN, 2016).

A ferramenta que apresentou mais resultados foi o Pfam, com 57 predições, seguido pelo SMART com 51 resultados. O CATCH foi a ferramenta com menor quantidade de resultados, apenas 15 predições.

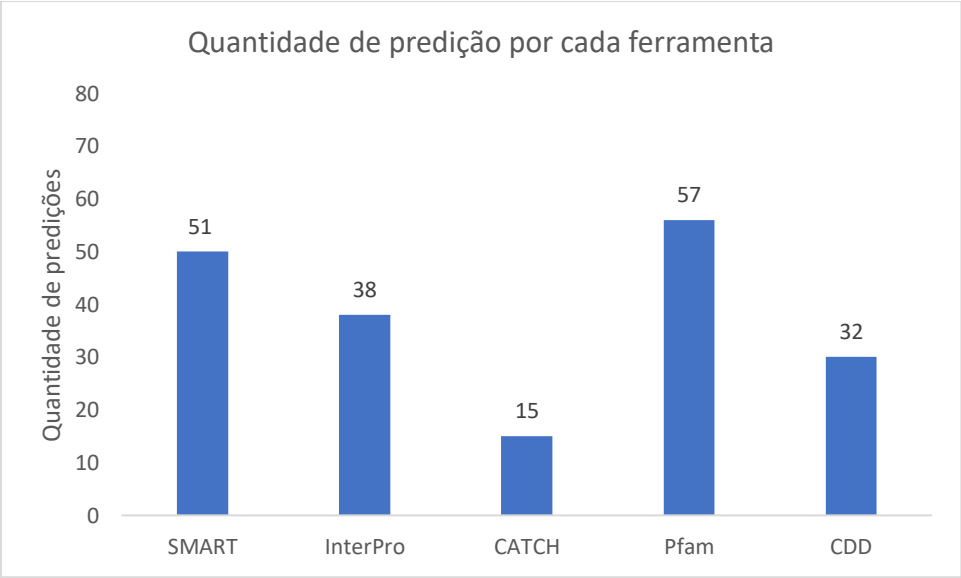

Gráfico 1. Número de proteínas hipotéticas (HPs) previstas por cada ferramenta.

Após a busca por domínios e famílias, foi construído um diagrama de Venn para observar similaridades entre os resultados e definir, assim, um melhor critério para a seleção de proteínas de alta confiabilidade que teriam sua anotação otimizadas.

De acordo com o diagrama, demonstrado na Figura 2, um total de 7 proteínas obtiveram resultados similares em todas das cinco ferramentas utilizadas. Foram elas: FWK01_00050, FWK01_02830, FWK01_03355, FWK01_08665, FWK01_18275, FWK01_27080 e FWK01_29045.

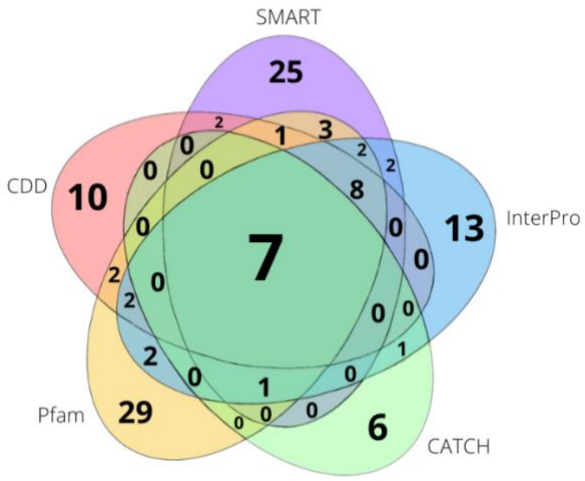

Figura 2. Diagrama de Venn para demonstração de similaridade entre os resultados das predições de domínios e superfamílias. Os números mais externos indicam a quantidade de resultados exclusivos de cada preditor.

A partir da observação da similaridade entre os resultados obtidos, foi definido um critério para selecionar as proteínas de alta confiabilidade, conforme demonstrado na Figura 4. Ao final, foram selecionadas 16 proteínas de alta confiabilidade com resultados similares em ao menos quatro das ferramentas usadas para predição.

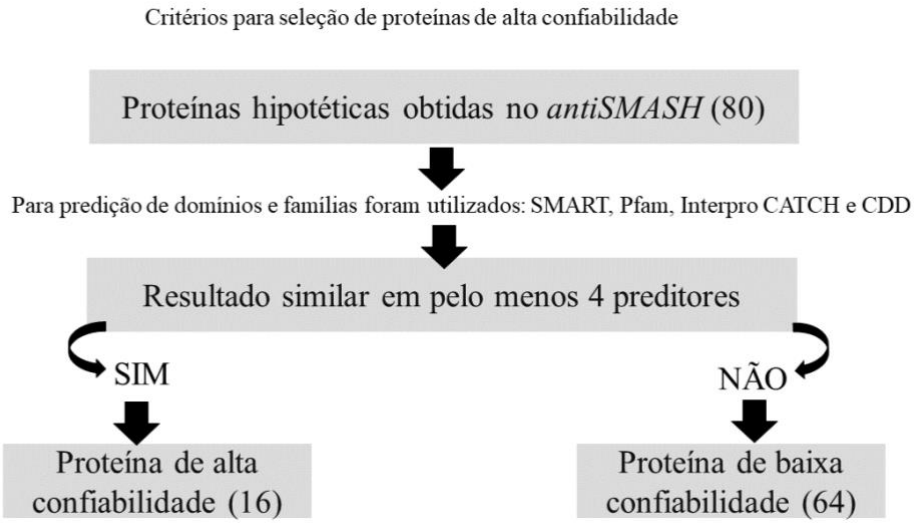

Figura 3. Critérios de confiabilidade para seleção de proteínas de alta confiabilidade.

Ademais, foi possível observar também foi possível observar a posição das proteínas hipotéticas de alta confiabilidade no genoma da *Pantanalinema* sp. GBBB05. Pode-se inferir a proximidade entre os agrupamentos gênicos de interesse através da posição das proteínas no genoma.

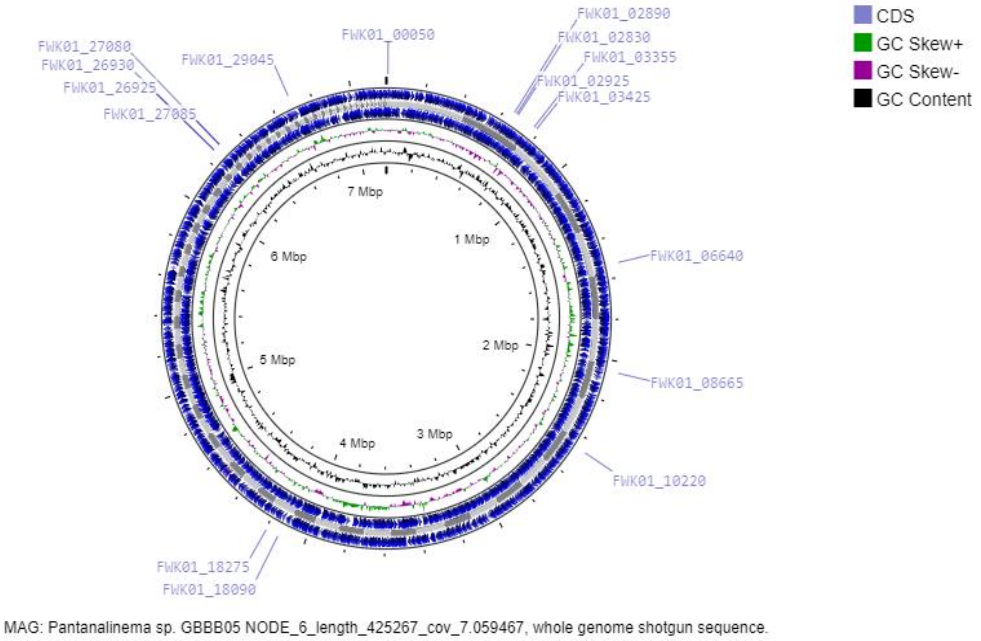

Figura 4. Mapa genômico da *Pantanalinema* sp. GBBB05 feito através do CG View, demonstrando a localização das proteínas hipotéticas de alta confiabilidade.

As proteínas de alta confiabilidade tiveram suas sequências de aminoácidos colocadas no banco de dados UniProt, o qual gerou um número de acesso que foi pesquisado no banco de dados Interpro, definindo uma possível função para as HCs.

Tabela 2. Classificação funcional de domínios para as HCs-HPs utilizando o Uniprot e Interpro.

ID da HCp	Agrupamento	Entrada Uniprot	Classificação interpro
FWK01_00050	1.1 Terpeno	A0A2T1DX45_9CYAN	Peptidase C39, processamento de bacteriocina
FWK01_02830	2.1 Betalactona	K8GQ85_9CYAN	Ficobilissoma, subunidade alfa/beta
FWK01_02845	2.1 Betalactona	A0A2T1EGM5	Domínio NnrU
FWK01_02890	2.1 Betalactona	K8GJJ2_9CYAN	Proteína YlqD
FWK01_02925	2.1 Betalactona	A0A2T1ENK8_9CYAN	Enzima de biossíntese de homocisteína, incorporação de enxofre
FWK01_03355	3.1 Resorcinol	A0A0V7ZHQ0_9CYAN	Domínio nucleosídio fosforilase
FWK01_03425	3.1 Resorcinol	A0A1Z4HNH5_9NOSO	Porina Seletiva de Carboidratos OprB
FWK01_08665	8.1 NRPS	A0A1E5QJN4_9CYAN	Domínio de histidina quinase
FWK01_10220	10.1 NRPS	A0A1Q8Z6K0_9CYAN	Porina Seletiva de Carboidratos OprB

FWK01_18090	21.1 NRPS, NRPS-like	A0A2T1CEP9_9CYAN	2A/B/C tipo peroxirredoxina
FWK01_18275	21.1 NRPS, NRPS-like	A0A367QUI3_9NOSO	Glicosil transferase, família 28, C-terminal
FWK01_26925	49.1 Terpeno	K9VZF6_9CYAN	Superfamília de domínio contendo repetição WD40
FWK01_26930	49.1 Terpeno	K9VZF6_9CYAN	Superfamília de domínio contendo repetição WD40
FWK01_27080	50.1 Lantipeptidio-classe-v	A0A1U7I3B5_9CYAN	Superfamília do tipo PGBD/ Tipo de ligação de peptidoglicano
FWK01_27085	50.1 Lantipeptidio-classe-v	A0A2T1ER10_9CYAN	Porina Seletiva de Carboidratos OprB
FWK01_29045	62.1 NRPS	B2IV05_NOSP7	Citocromo b561, bacteriano/ Ni-hidrogenase

As proteínas FWK01_02845, FWK01_02890 e FWK01_02925 foram caracterizadas com funções que ainda não são muito claras. A sequência da proteína FWK01_02845 demonstrou um percentual de identidade de 78, 63% e 80,17% com a proteína da família NnrU das cianobactérias *Leptolyngbya* sp. O-77 e *Crinalium epipsammum*, respectivamente. Pressupõe-se que proteínas com domínio NnrU estão envolvidas na oxirredução do ácido nítrico (CHEN; LI; WURTZEL, 2010).

As proteínas FWK01_03425, FWK01_10220 e FWK01_27085 foram caracterizadas como pertencentes à família porina seletiva de carboidrato OprB. As proteínas dessas famílias estão envolvidas na difusão de carboidratos através da membrana (WYLIE; WOROBEC, 1995).

A sequência de aminoácidos da proteína FWK01_00050 demonstrou um percentual de identidade de 69,39% com a sequência da proteína de *Stenomitos frigidus*, caracterizada como uma proteína da família cisteina peptidase C39. Essa família de proteínas é composta por proteases de maturação para bacteriocinas peptídicas. As bacteriocinas são peptídeos antimicrobianos com potencial aplicação na indústria alimentícia, podendo ser utilizadas como bioconservantes alimentícios e uma fonte alternativa aos antibióticos (OGAKI; FURLANETO; MAIA, 2015). A proteína FWK01_02830 foi predita como uma proteína que atua na formação dos ficobilossomos. Os ficobilossomos são complexos proteícos formados por filobiliproteínas, constituindo uma estrutura primordial para a captação de luz, desempenhando uma função fotossintetizante. Quando isoladas e purificadas, nota-se que essas proteínas possuem subunidades (alfa, beta e gama). Referindo-se às aplicações dessa molécula, elas já foram utilizadas como corantes e estabilizantes de alimentos, agentes anti-inflamatórios, antioxidantes e marcadores de biomoléculas (NOBRE, 2015).

A peroxirredoxina é uma proteína com grande capacidade anti-inflamatória e antioxidante que vem sendo muito estudada por ser capaz de inibir a proliferação e a glicólise no câncer gástrico (GUO et al., 2015; ZHANG et al., 2018). A proteína FWK01_18090 foi caracterizada como uma peroxirredoxina e apresentou 79,55% de identidade com a proteína da cianobactéria *Kovacikia minuta* caracterizada como uma proteína da família AhpC/TSA que possui proteínas relacionadas com a redução de aquil hidroperóxido (AhpC) e também contém enzimas antioxidantes tiorredoxina redutase (TSA) (CHAE et al., 1994).

5.1.2 Parâmetros físico-químicos

Por meio do *Expasy Protparam*, as sequências das proteínas de alta confiabilidade foram submetidas para aferição dos parâmetros físico químicos. Através dessa ferramenta, foi possível calcular o número de aminoácidos, peso molecular, ponto isoelétrico, coeficiente de extinção, índice alifático, valor GRAVY e índice de estabilidade. O ponto isoelétrico (pi) corresponde ao valor de pH no qual as cargas positivas e negativas de uma molécula se igualam, ou seja, se anulam. O pi varia para cada proteína, pois dependerá da polaridade dos radicais dos aminoácidos que a constituem (TRINDADE et al., 2012). Através do valor do índice alifático, é possível determinar o volume que as cadeias alifáticas ocupam (alanina, valina, isoleucina e leucina), o que irá implicar na estabilidade térmica. Quanto maior é o índice alifático, maior é a estabilidade térmica da proteína. Já o índice de estabilidade permite a classificação da proteína em estável ou instável, geralmente valores menores que 40 podem indicar uma proteína estável (RIBEIRO; BRANCO; CHOUPINA, 2021; THAKUR et al., 2020).

O índice de hidropraticidade média, valor GRAVY (-sigla em inglês), para um peptídeo ou proteína é obtido a partir do cálculo da soma dos valores de hidropatia de todos os aminoácidos e dividido pela quantidade total de resíduos na sequência (BEZERRA; QUEIROZ; FREIRE, 2018). Essa média irá indicar a interação da proteína com a água, valores negativos indicam proteínas hidrofílicas, e quanto maior o valor, maior será o índice de hidrofobicidade da proteína (RIBEIRO; BRANCO; CHOUPINA, 2021).

Com exceção das proteínas FWK01_00050, FWK01_02845 e FWK01_29045, que obtiveram valores positivos, todas as demais proteínas são hidrofóbicas. Já em termos de estabilidade, 8 proteínas foram classificadas como estáveis e 8 como instáveis.

Tabela 3. Propriedades físico químicas das HCs-Hps.

ID da HCp	Nº A	PM	PI	CE	Índice de estabilidade calculado	Indice de estabilidade classificação	IA	GRAVY
FWK01_00050	358	39351.04	9.54	56045	36.33	estável	112.82	0.278
FWK01_02830	158	18171.00	5.14	13075	43.14	instável	103.16	-0. 236
FWK01_02845	236	26867.71	9.20	70930	29.07	estável	120.68	0. 571
FWK01_02890	149	17168.69	4.95	5500	55.50	instável	100.60	-0. 544
FWK01_02925	394	42968.22	6 .02	48150	37.08	estável	96.98	-0. 040
FWK01_03355	329	35817.97	5.56	33585	43.41	instável	96.11	-0. 161
FWK01_03425	532	57783.17	4.61	74050	22.42	estável	83.25	-0. 152
FWK01_08665	531	59025,67	5,3	53205	43,97	instável	108,4	-0, 020
FWK01_10220	585	62789,3	4,81	60405	27,68	estável	85,11	-0,118
FWK01_18090	264	29132,45	6,58	41160	40,49	instável	94,66	-0,047
FWK01_18275	388	43938,53	6,42	61670	41,97	instável	96,78	-0,179
FWK01_26925	422	47443,78	5,97	81610	45,57	instável	91,28	-0,275
FWK01_26930	314	33885,38	6,11	47900	45,59	instável	90,99	-0, 075
FWK01_27080	277	30504,79	5,5	41035	30,17	estável	90,94	-0, 166
FWK01_27085	543	58257,62	4,59	67965	24,44	estável	82,36	-0, 166
FWK01_29045	110	12821,28	11,84	26470	38,59	estável	117,91	0,42

Legenda: Aa, aminoácidos; PM, peso molecular; PI, ponto isoelético; CE, coeficiente de extinção; IA índice alifático; GRAVY, *Grand Average of hydrophathicity* (índice de hidropaticidade média).

5.1.3 Determinação de Localização subcelular

A determinação da localização subcelular ajuda na compreensão do papel da proteína na célula. A análise de localização subcelular de todas as proteínas hipotéticas pelo CELLO, PSORT e PSLpred demonstrou predominância para localização citoplasmática (60%), 11% das proteínas são de membrana, 21% de função desconhecida e 8% das proteínas foram preditas com localização extracelular.

Em relação às proteínas de alta confiabilidade, as proteínas FWK01_00050, FWK01_02845, FWK01_03425, FWK01_08665, FWK01_10220 e FWK01_29045 foram caracterizadas como proteínas de membrana. As proteínas FWK01_02830, FWK01_02890, FWK01_02925, FWK01_03355 e FWK01_27080 apresentaram localização subcelular citoplasmática.

Não foi possível definir a localização subcelular das proteínas FWK01_18275, FWK01_18090, FWK01_26930, e FWK01_27085 pois não apresentaram convergência de resultados. Nenhuma das proteínas de alta confiabilidade foi caracterizada como proteína extracelular.

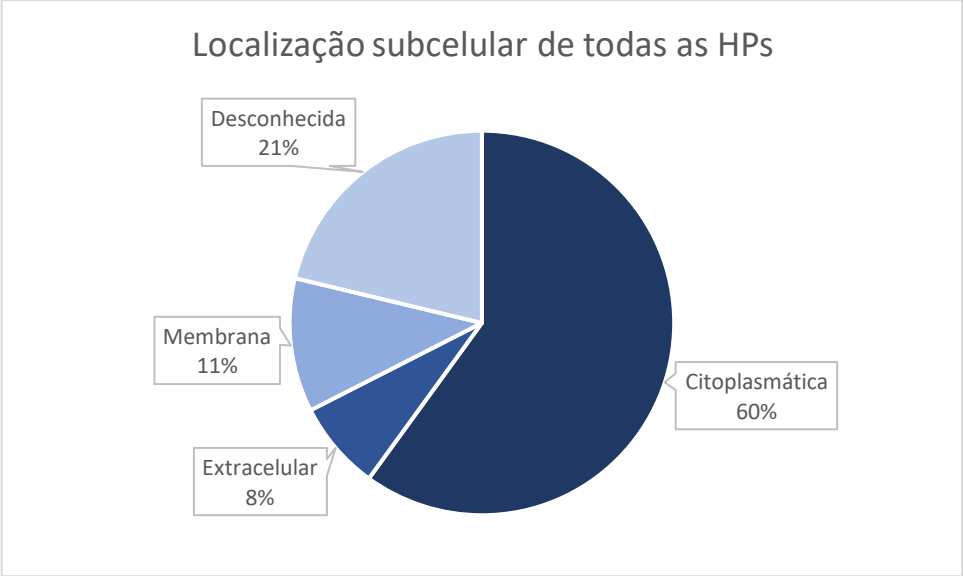

Gráfico 2. Localização subcelular de todas as HPs, utilizando as ferramentas PSLPred, CELLO e PsortB.

A análise do secretoma é de extrema importância para que sejam compreendidos os processos extracelulares (MATAFORA; BACHI, 2020).

Através de sistemas de secreção, as bactérias podem secretar proteínas e DNA para o meio extracelular e para dentro de células (procarióticas e eucarióticas). É através desses sistemas que bactérias patogênicas, por exemplo, conseguem evadir as defesas de hospedeiros, além de estar envolvido na competição entre células procarióticas (ALMEIDA, 2016).

A análise realizada pelo SecretomeP para predição de proteínas com atividade secretora demonstrou que somente as proteínas FWK01_02845, FWK01_10220, FWK01_26930 e FWK01_27085 são proteínas secretadas, correspondendo a 25% das proteínas de alta confiabilidade.

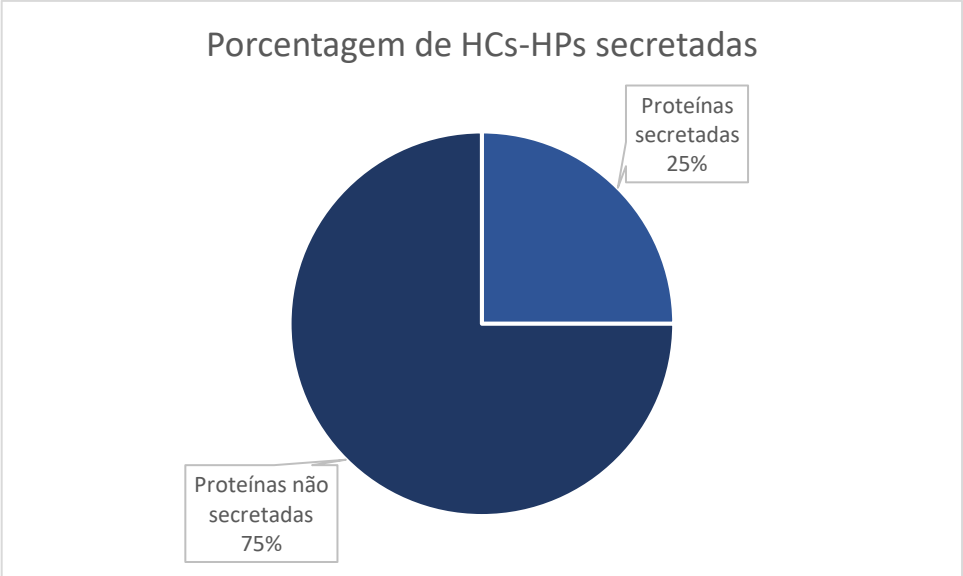

Gráfico 3. Gráfico com percentual de proteínas secretadas identificadas pelo Secretomep 2.0.

Já análise feita no SignalP busca identificar peptídeos sinais de sequências de aminoácidos. Proteínas com esse tipo de sinal são direcionadas para a via secretora mas nem sempre são secretadas (ALMAGRO ARMENTEROS et al., 2019). O resultado dessa análise previu apenas as proteínas FWK01_02845 e FWK01_27085 com peptídeo sinal.

5.1.4 Hélice transmembrana e previsão de topologia

As proteínas transmembranas são aquelas que atravessam a bicamada lipídica da membrana plasmática (HALLGREN et al., 2022). Com a ajuda do *Deep*TMHMM e CCTOP foi possível prever a existência e o número de hélices transmembrana das proteínas analisadas. O SOSUI foi utilizado para identificar proteínas de membranas e proteínas solúveis. Trabalhos que buscam identificar proteínas de membrana e fazer uma previsão de topologia das proteínas são de grande importância para a biologia celular pois essas proteínas estão envolvidas em vários processos os quais atuam na comunicação entre o meio intracelular e extracelular, como o processo de transporte de íons e soluto através da membrana e produção de energia (DA CRUZ; OTAKE; DELGADO-CAÑEDO, 2013; DOBSON; REMÉNYI; TUSNÁDY, 2015).

Das 16 proteínas de alta confiabilidade somente as proteínas FWK01_00050, FWK01_02845, FWK01_08665 e FWK01_29045 foram identificadas como proteínas de membrana.

Tabela 4. Predição de Hélices transmembrana *Deep*TMHMM e CCTOP, e caracterização de solubilidade pelo SOSUI.

HC-Hp	Agrupamento	SOSUI	CCTOP	TMHMM
FWK01_00050	1.1 Terpeno	PROTEÍNA DE MEMBRANA	6	4
FWK01_02830	2.1 Betalactona	PROTEÍNA SOLÚVEL	1	1
FWK01_02845	2.1 Betalactona	PROTEÍNA DE MEMBRANA	4	5
FWK01_02890	2.1 Betalactona	PROTEÍNA SOLÚVEL	0	0
FWK01_02925	2.1 Betalactona	PROTEÍNA SOLÚVEL	2	0
FWK01_03355	3.1 Resorcinol	PROTEÍNA SOLÚVEL	3	0
FWK01_03425	3.1 Resorcinol	PROTEÍNA SOLÚVEL	1	0
FWK01_08665	8.1 NRPS	PROTEÍNA DE MEMBRANA	1	1
FWK01_10220	10.1 NRPS	PROTEÍNA SOLÚVEL	1	1
FWK01_18090	21.1 NRPS, NRPS-like	PROTEÍNA SOLÚVEL	0	0
FWK01_18275	21.1 NRPS, NRPS-like	PROTEÍNA SOLÚVEL	1	0
FWK01_26925	49.1 Terpeno	PROTEÍNA SOLÚVEL	1	0
FWK01_26930	49.1 Terpeno	PROTEÍNA SOLÚVEL	1	0
FWK01_27080	50.1 Lantipeptidio-classe-v	PROTEÍNA SOLÚVEL	1	0
FWK01_27085	50.1 Lantipeptidio-classe-v	PROTEÍNA SOLÚVEL	6	0
FWK01_29045	62.1 NRPS	PROTEÍNA DE MEMBRANA	2	2

5.2 Avaliação de desempenho

A sensibilidade de um teste é definida como a taxa de identificação de verdadeiro positivos, enquanto a especifidade, trata-se da taxa de verdadeiros negativos. A acurácia, por sua vez, corresponde à taxa de acerto do modelo (POLO; MIOT, 2020).

O CDD foi o preditor com maior taxa de sensibilidade (84,4%). O Interpro teve a maior porcentagem de acurácia (86,3%) e o Pfam foi a ferramenta que demonstrou uma maior taxa de especificidade (91,7%).

Tabela 5. Resultado da análise de curva ROC realizada na plataforma *Online Curve ROC*.

SOFTWARE	SENSIBILIDADE	ESPECIFICIDADE	ACURÁCIA	ROC ÁREA
SMART	62.5%	84.4%	71.3%	0.8020
INTERPRO	81.6%	88.1%	85%	0.9447
CATCH	83.3%	66.2%	68.8%	0.8474
PFAM	64.3%	91.7%	72.5%	0.8855
CDD	84.4%	81.3%	82.5%	0.9031
MÉDIA	**75,22%**	**82,34%**	**76,02%**	**0.87654**

5.3 Filogenia Molecular e filogenômica

O proteoma da *Pantanalinema* sp. GBBB05 foi comparado com os proteomas de outras 19 linhagens, algumas de gêneros pertencentes da mesma família da *Pantalinema* e outras de família diferentes como a *Geitlerinema* e *Oscillatoria* da família *Oscilatoriaceae*. A análise feita no TYGS por meio do FastME 2.1.6.1, confirmou a identificação da GBBB05, como membro da família *Leptolyngbyaceae* ao situá-la próxima de duas espécies do gênero *Leptolyngbya*.

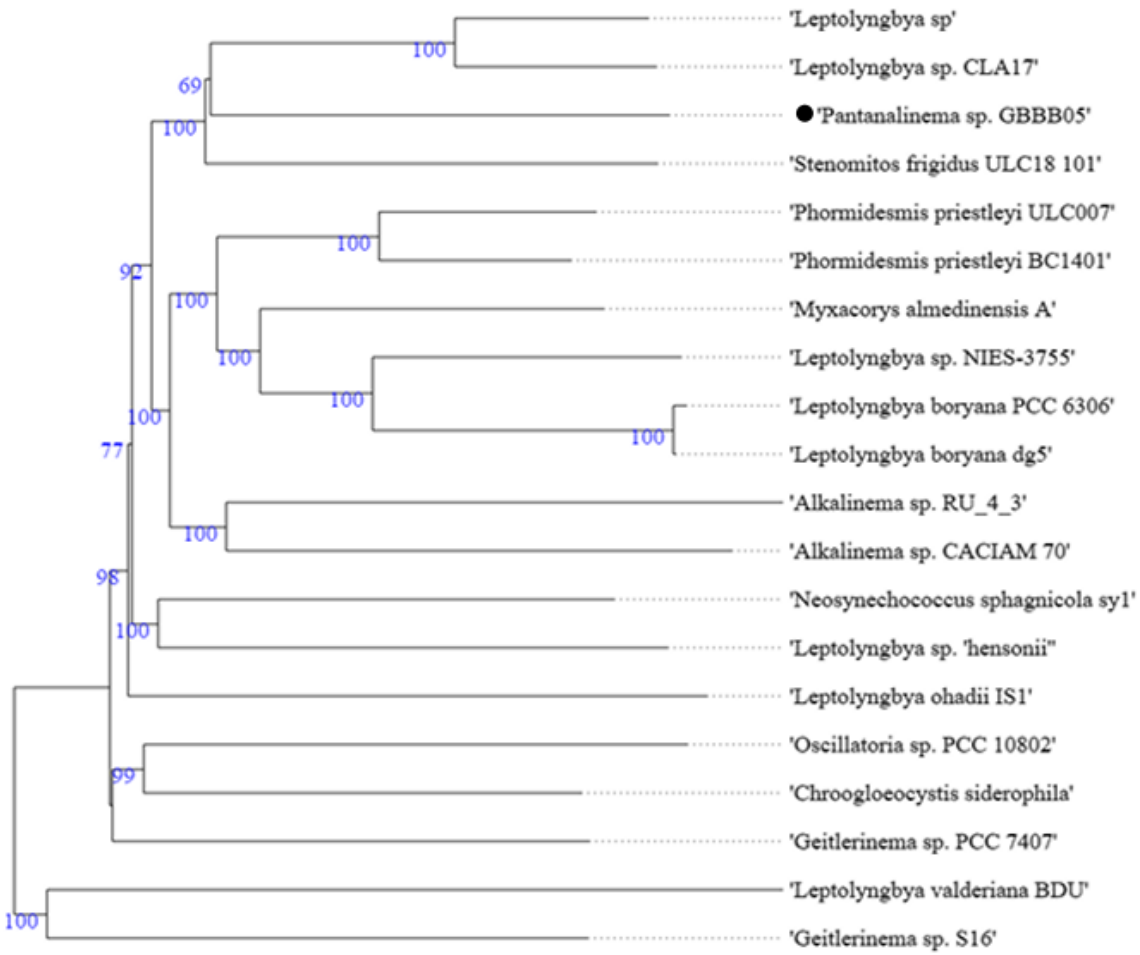

Figura 5. A árvore foi inferida com FastME 2.1.6.1 a partir de distâncias GBDP baseadas no proteoma inteiro. Os comprimentos dos ramos são dimensionados através da fórmula de distância GBDP *d*5. Os valores de ramificação são valores de suporte de pseudo-bootstrap GBDP > 60% de 100 replicações, com um suporte médio de ramificação de 93,6%. A árvore foi enraizada no ponto médio.

Foram construídas 6 árvores gênicas com base na sequência das proteínas codificadas por genes sintéticos centrais.

A proteína citramalato síntase (FWK01_02875) do agrupamento 2.1 betalactona é codificada pelo gene *cimA* e é descrita como uma proteína envolvida com uma das etapas da

biossíntese da enzima isoleucina, que por sua vez é responsável pela biossíntese de aminoácidos (KELLY et al., 2014; SHANER, 1997).

A árvore gerada a partir da sequência de aminoácidos dessa proteína mostrou que a proteína citramalato síntase da GBBB05 possui mais homologia com a mesma proteína de duas espécies da família *Leplyngbyaceae,* que é a mesma família a qual o gênero *Pantanalinema* pertence. A sequência de aminoácidos da GBBB05 apresentou um percentual de identidade de 82,49% com a sequência da cianobactéria da *Leptothermofonsia sichuanensis* e 80,75% com a proteína da *Leptolyngbyaceae cyanobacterium HOT.MB2.61*.

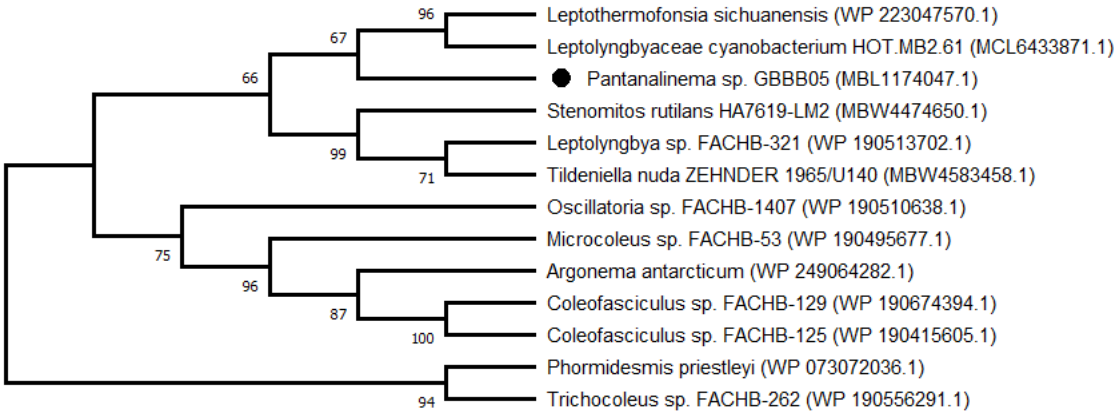

Figura 6. A foi inferida usando o método *Neighbor-Joining* e as distâncias evolutivas foram calculadas usando o método de correção de Poisson. Foi empregado o teste de bootstrap com 1000 replicatas, a porcentagem de árvores replicadas nas quais os táxons associados agrupados no teste de bootstrap são mostrados ao lado dos ramos. Os ramos correspondentes a partições reproduzidas em menos de 50% de réplicas de bootstrap são recolhidos. Esta análise envolveu 13 sequências de aminoácidos. As análises evolutivas foram realizadas no MEGA X. Em parênteses estão indicados os IDs das proteínas.

A proteína beta-cetoacil-ACP sintase III (FWK01_03380) do agrupamento 3.1 NRPS é descrita como responsável pela síntase de ácidos graxos (ALMEIDA-AMARAL et.al, 2014). Baseada na construção da árvore gênica que utilizou a sequência de aminoácidos dessa proteína, notou-se que a *Pantanalinema* sp. GBBB05 foi agrupada no mesmo clado de outra espécie da família *Leptolynbyaceae*, a *Leptothermofonsia sichuanensis*. O percentual de identidade entre as duas sequências foi de 79,35%.

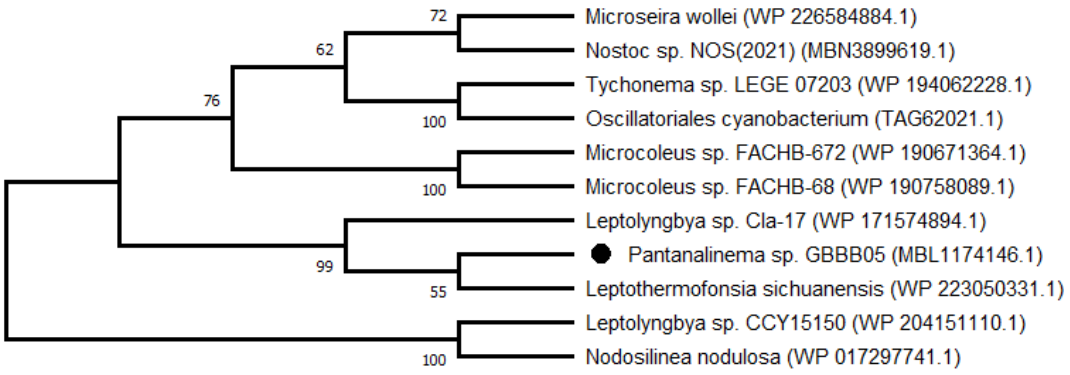

Figura 7. A foi inferida usando o método *Neighbor-Joining* e as distâncias evolutivas foram calculadas usando o método de correção de Poisson. Foi empregado o teste de bootstrap com 1000 replicatas, a porcentagem de árvores replicadas nas quais os táxons associados agrupados no teste de bootstrap são mostrados ao lado dos ramos. Os ramos correspondentes a partições reproduzidas em menos de 50% de réplicas de bootstrap são recolhidos. Esta análise envolveu 11 sequências de aminoácidos. As análises evolutivas foram realizadas no MEGA X. Em parênteses estão indicados os IDs das proteínas.

A sequência de aminoácidos das proteína contendo domínio de adenilação de aminoácidos dos agrupamentos 5.1 NRPS, T1PKS (FWK01_06765); 10.1 NRPS (FWK01_10200); 21 NRPS-NRPS,*like* (FWK01_18155); e 62.1 NRPS (FWK01_28975) foram utilizadas para gerar as demais árvores gênicas. Os resultados das árvores gênicas construídas a partir das proteínas dos agrupamentos 5.1 e 62.1 sugerem que a GBBB05 tenha recebido os genes para as respectivas proteínas por transferência horizontal. A transferência horizontal de genes é um mecanismo capaz de aumentar a variabilidade genética de organismos procariotos, consiste na troca de matéria genético entre espécies não relacionadas, que em procariotos pode ocorrer através da transformação, conjugação e transdução (BROWN, 2003)**.**

Na árvore gerada pela sequência de aminoácidos da proteína FWK01_06765 do agrupamento 5.1, foi demonstrada uma homologia entre as sequências da GBBB05 e da cianobactéria da classe *Oscillatoriophycideae,* com um valor de bootstrap de 59%. Ambas as proteínas possuem domínio de adenilação de aminoácidos e têm percentual de identidade de 72,67%.

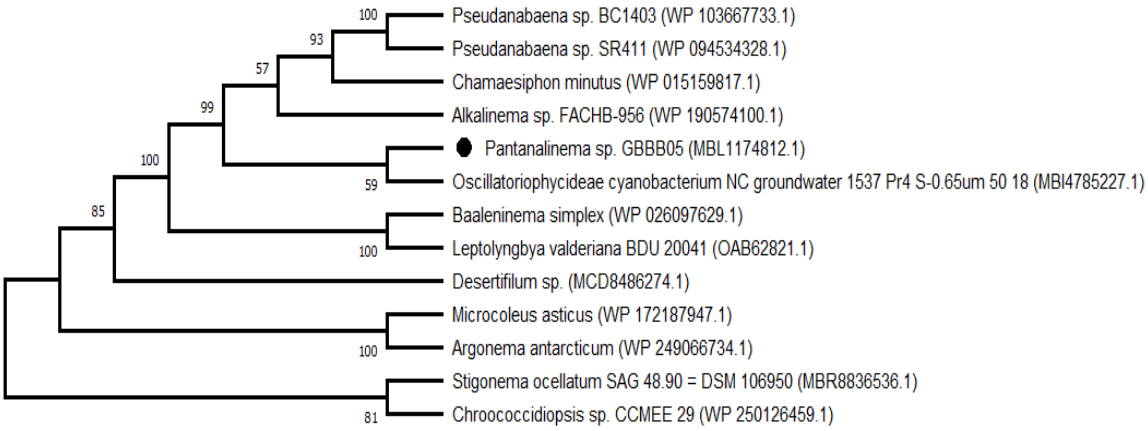

Figura 8. A foi inferida usando o método *Neighbor-Joining* e as distâncias evolutivas foram calculadas usando o método de correção de Poisson. Foi empregado o teste de bootstrap com 1000 replicatas, a porcentagem de árvores replicadas nas quais os táxons associados agrupados no teste de bootstrap são mostrados ao lado dos ramos. Os ramos correspondentes a partições reproduzidas em menos de 50% de réplicas de bootstrap são recolhidos. Esta análise envolveu 13 sequências de aminoácidos. As análises evolutivas foram realizadas no MEGA X. Em parênteses estão indicados os IDs das proteínas.

A árvore gênica construída com a proteína FWK01_10200 do agrupamento 10.1 demonstrou homologia entre a GBBB05, uma espécie do gênero *Leptolyngbya* (a qual pertence à mesma família da *Pantanalinema*), e uma espécie do gênero *Phormidium* com *bootstrap* de 100%. A proteína da *Leptolyngbya* sp. FACHB-321, que apresentou um percentual de identidade de 68,69% com a proteína da GBBB05, contêm domínio de síntese de peptídeos e domínio PiLZ, o qual faz parte da proteína de ligação bacteriana c-di-GMP, que por sua vez atua como um mensageiro secundário universal envolvida na regulação de alguns processos como produção de polissacarídeos extracelulares, formação de motilidade e outros comportamentos celulares em diversas bactérias (AMIKAM; GALPERIN, 2006; PAUL et al., 2010) .

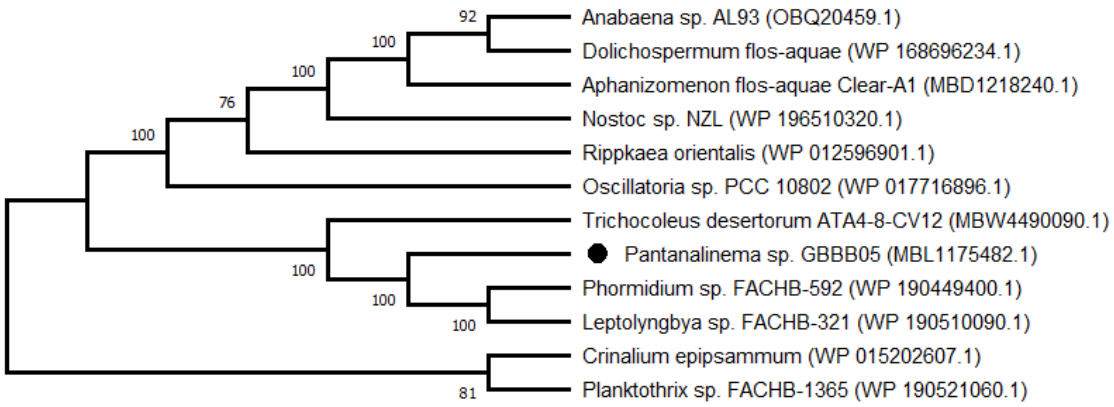

Figura 9. A foi inferida usando o método *Neighbor-Joining* e as distâncias evolutivas foram calculadas usando o método de correção de Poisson. Foi empregado o teste de bootstrap com 1000 replicatas, a porcentagem de árvores replicadas nas quais os táxons associados agrupados no teste de bootstrap são mostrados ao lado dos ramos. Os ramos correspondentes a partições reproduzidas em menos de 50% de réplicas de bootstrap são recolhidos. Esta análise envolveu 12 sequências de aminoácidos. As análises evolutivas foram realizadas no MEGA X. Em parênteses estão indicados os IDs das proteínas.

A proteína FWK01_18155 do agrupamento 21.1 NRPS demonstrou homologia com *bootstrap* de 100% com a proteína que contem domínio de adenilação de aminoácidos da cianobactéria JSC-12 da família *Leptolynbyaceae*, mesma família a qual o gênero *Pantanalinema* pertence. O percentual de identidade entre as duas sequências é de 77,42%.

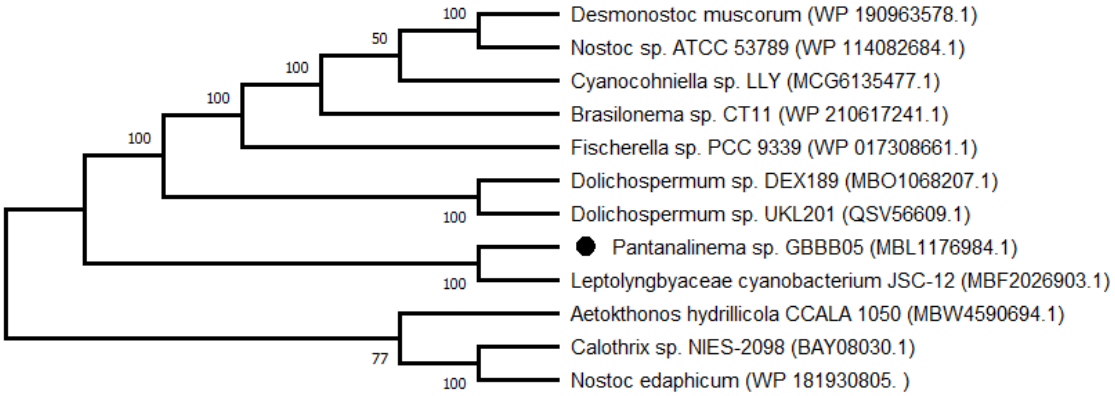

Figura 10. A foi inferida usando o método *Neighbor-Joining* e as distâncias evolutivas foram calculadas usando o método de correção de Poisson. Foi empregado o teste de bootstrap com 1000 replicatas, a porcentagem de árvores replicadas nas quais os táxons associados agrupados no teste de bootstrap são mostrados ao lado dos ramos. Os ramos correspondentes a partições reproduzidas em menos de 50% de réplicas de bootstrap são recolhidos. Esta análise envolveu 12 sequências de aminoácidos. As análises evolutivas foram realizadas no MEGA X. Em parênteses estão indicados os IDs das proteínas.

Na árvore gênica gerada utilizando a proteína do agrupamento 62.1, a *Pantanalinema* sp. GBBB05 foi agrupada no mesmo clado que duas espécies dos gêneros *Oscillatoria*, da família *Oscilatoriaceae*, e *Nodosilinea*, da família das *Proclorotricáceas*, com *bootstrap* de 97%. As sequências das proteínas das espécies da *Nodosilinea* e *Oscillatoria* possuem domínio pra síntese de peptídeos e domínio *metiltransferase*. Este último está envolvido com a transferência de metil utilizando o substrato S-adenosil-L-metionina, criando o produto S-adenosil-L-homocisteína (MARTIN, 2002; SCHUBERT; BLUMENTHAL; CHENG, 2003).

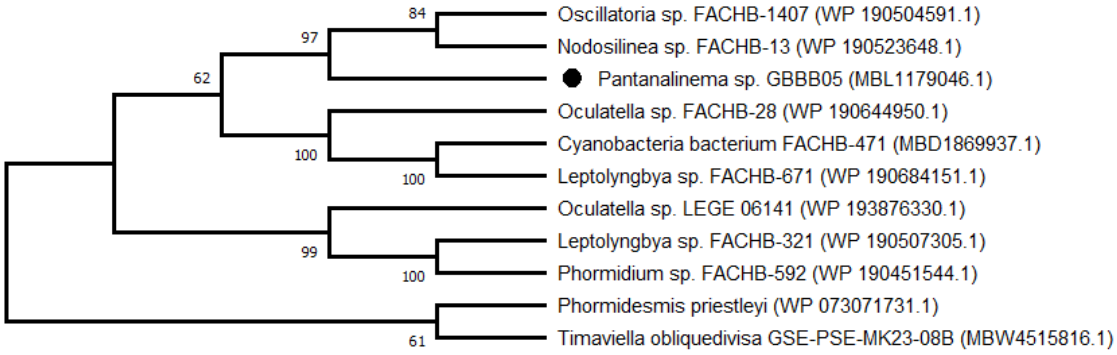

Figura 11. A foi inferida usando o método *Neighbor-Joining* e as distâncias evolutivas foram calculadas usando o método de correção de Poisson. Foi empregado o teste de bootstrap com 1000 replicatas, a porcentagem de árvores replicadas nas quais os táxons associados agrupados no teste de bootstrap são mostrados ao lado dos ramos. Os ramos correspondentes a partições reproduzidas em menos de 50% de réplicas de bootstrap são recolhidos. Esta análise envolveu 11sequências de aminoácidos. As análises evolutivas foram realizadas no MEGA X. Em parênteses estão indicados os IDs das proteínas.

5.4 Análise de sintenia

Foram realizadas análises de sintenia de dois agrupamentos da *Pantanalinema* sp. GBBB05: O agrupamento 5.1 NRPS-T1PKS; e o agrupamento 10.1 NRPS.

A sintenia do agrupamento 5.1 NRPS-T1PKS da GBBB05 foi feita com os agrupamentos NRPS, NRPS-like e NRPS-T1PKS das espécies PCC 9339 e PCC 9431, respectivamente, do gênero *Fischerella,* ambos indicados pelo *antiSMASH*. O agrupamento da espécie PCC 9339 está envolvido com a biossíntese de aranozol, enquanto o da espécie PCC 9431 está envolvido com a biossíntese de halogenases dependentes de ferro (II)/α-cetoglutarato (MOOSMANN et al., 2018). A análise demonstrou uma relação de sintenia pouco significativa entre esses agrupamentos e o agrupamento da espécie da *Pantanalinema*. A GBBB05 compartilha com a espécie PCC 9431 as proteínas de domínio PKS_KS com 36% de identidade e a proteína de domínio de heterociclização com 44% de identidade.

Já com a espécie PCC 9339 a GBBB05 compartilha além da proteína de domínio PKS_KS, a proteína de Domínio de Condensação LCL (Domínio de condensação ligando um L-aminoácido a um peptídeo que termina com um L-aminoácido) (RAUSCH et al. 2007). É possível observar uma inversão de sentido dessas proteínas no agrupamento da espécie da *Fischerella*.

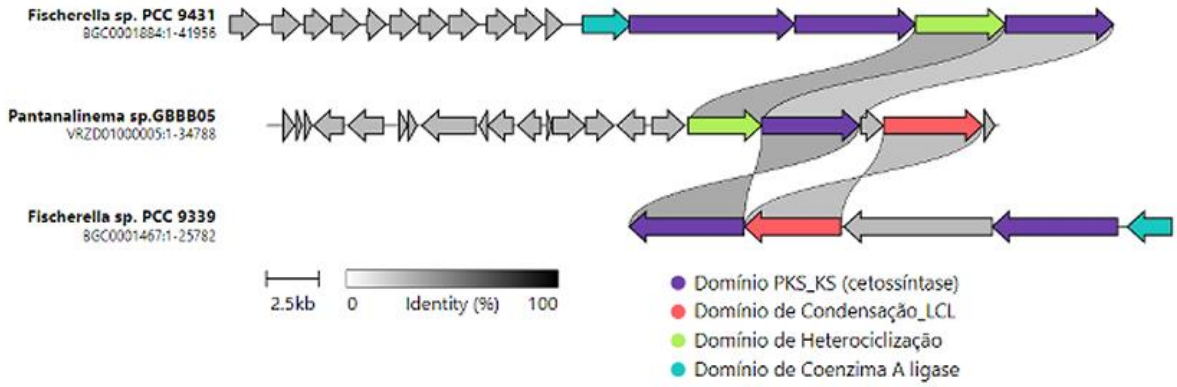

Figura 12. Análise de sintenia do grupamento 5.1 NRPS- T1PKS da *Pantanalinema* sp. GBBB05 com os agrupamentos das cianobatérias *Fischerella* sp. PCCC 9431 e *Fischerella* sp. PCCC 9339 indicados pelo *antiSMASH* utilizando *clinker* e *clustermap.js.*

O agrupamento 5.1 NRPS-T1PKS também foi comparado com o agrupamento NRPS-T1PKS da cianobactéria *Oscillatoriophycideae cyanobacterium* NC_groundwater_1537_Pr4_S-0.65um_50_18. A comparação mostrou identidade significativa entre 10 proteínas dos agrupamentos e também similaridade quanto à organização destas proteínas nas respectivas vias metabólicas.

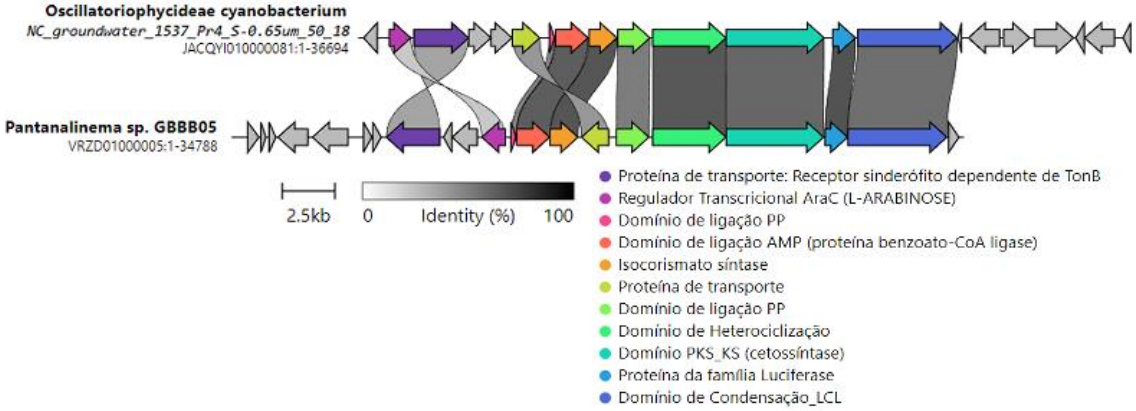

Figura 13. Análise de sintenia entre os agrupamentos NRPS-T1PKS da *Pantanalinema* sp. GBBB05 e da espécie de cianobactéria da classe *Oscillatoriophycideae* utilizando *clinker* e *clustermap.js.*

O agrupamento 10.1 NRPS foi comparado com os agrupamentos: NRPS-T1PKS da *Nostoc* sp. GSV224 e o agrupamento NRPS da *Nostoc* sp. ATCC 53789, ambos envolvidos na biossíntese de nostopeptólido A2; e com o agrupamento NRPS da espécie *Planktothrix agardhii* NIVA-CYA 116, responsável pela produção de cianopeptolina. A análise demonstrou pouca semelhança entre os agrupamentos indicados pelo *antiSMASH* e o agrupamento da

Pantanalinema sp. GBBB05, que compartilha somente a proteína de domínio de condensação com um baixo percentual de identidade com os demais agrupamentos.

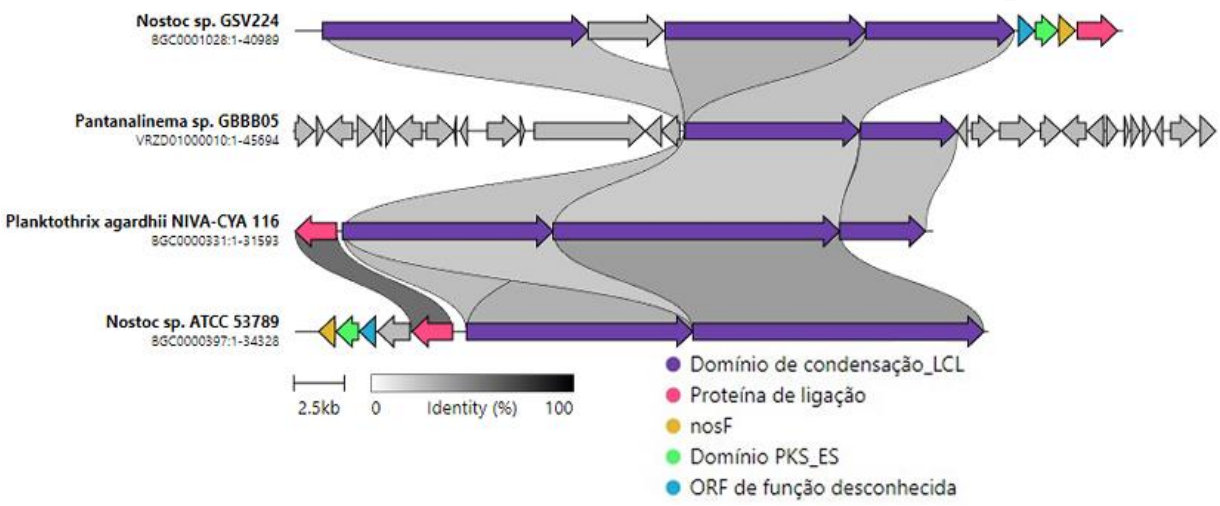

Figura 14. Análise de sintenia do agrupamento 10.1 NRPS da *Pantanalinema* sp. GBBB05 com os agrupamentos gênicos das espécies *Nostoc* sp. GSV224, *Nostoc* sp. ATCC 53789 e *Planktothrix agardhii* indicados pelo *antiSMASH* utilizando *clinker* e *clustermap.js*.

Além da comparação com os agrupamentos indicados pelo *antiSMASH*, o agrupamento 10.1 NRPS também foi comparado com o agrupamento NRPS da cianobactéria *Trichocoleus desertorum* ATA4-8-CV12. As ciabactérias da espécie *desertorum* foram identificadas em regiões de deserto e pertencem a família *Trichocoleusaceae* (MUHLSTEINOVA et al., 2014).

A sintenia entre esses dois agrupamentos demonstrou relação de identidade significativa entre algumas proteínas e também uma inversão da ordem de como algumas das proteínas estão inseridas nessas vias metabólicas.

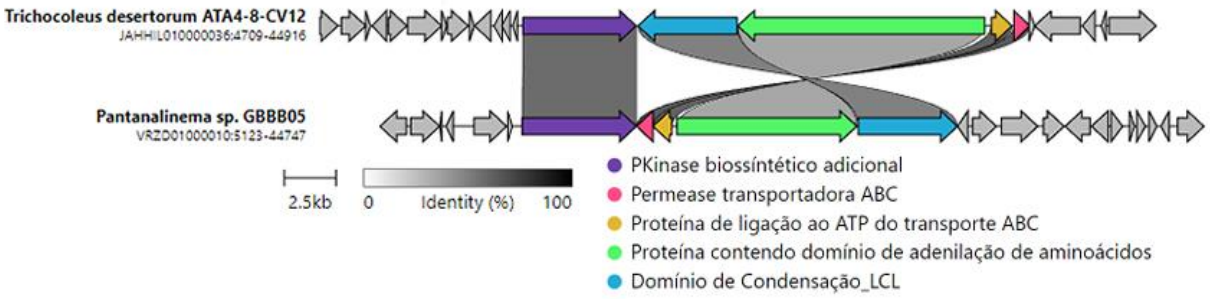

Figura 15. Análise de sintenia entre os agrupamentos NRPS da *Pantanalinema* sp. GBBB05 e a cepa ATA4-8-CV12 de espécie *Trichocoleus desertorum* utilizando *clinker* e *clustermap.js.*

6. CONSIDERAÇÕES FINAIS

Por meio da anotação manual das proteínas hipotéticas da *Pantanalinema* sp. GBBB05, foi possível identificar e caracterizar, com alta confiabilidade, 16 proteínas hipotéticas, lhes atribuindo funções distintas, além de fornecer informações quanto à localização subcelular e parâmetros físico-químicos, enfatizando a importância da bioinformática como meio para realizar estudos funcionais preliminares *in silico* que possam servir como base para futuras pesquisas *in vivo*. A espécie apresentou genes que possuem potencial para a biomedicina, biotecnologia e relevância industrial.

Além disso, este trabalho contribui para o estudo e a caracterização da biodiversidade local, uma vez que a espécie estudada foi identificada no Cerrado maranhense, demonstrando que tem grande potencial de abrigar uma ampla diversidade de cianobactérias. Enriquecendo também as informações ecológicas sobre o Parque Nacional da Chapada das Mesas que podem reforçar a importância da preservação ecológica do mesmo. Sendo assim faz-se necessário incentivar projetos de pesquisas que objetivem explorar essa biodiversidade, buscando identificar e caracterizar a microbiota local.

REFERÊNCIAS

ALMAGRO ARMENTEROS, J. J. et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. **Nature Biotechnology**, v. 37, n. 4, p. 420–423, 18 abr. 2019.

ALMEIDA, R. DAS N. Caracterização do papel do sistema de secreção do tipo VI bacteriano sobre a resposta imunológica inata de células de mamíferos infectados por Escherichia coli. 2016.

AMIKAM, D.; GALPERIN, M. Y. PilZ domain is part of the bacterial c-di-GMP binding protein. **Bioinformatics**, v. 22, n. 1, p. 3–6, 1 jan. 2006.

AZIZ, R. K. et al. The RAST Server: Rapid Annotations using Subsystems Technology. **BMC Genomics**, v. 9, n. 1, p. 75, 8 dez. 2008.

BATEMAN, A. et al. UniProt: the universal protein knowledgebase in 2021. **Nucleic Acids Research**, v. 49, n. D1, p. D480–D489, 8 jan. 2021.

BENSON, D. A. GenBank. **Nucleic Acids Research**, v. 28, n. 1, p. 15–18, 1 jan. 2000.

BEZERRA, L. C. C.; QUEIROZ, E. W. A. DE; FREIRE, J. E. DA C. PREDIÇÃO FÍSICO-QUÍMICA, MODELAGEM E ANÁLISE DO MECANISMO DE INTERAÇÃO DA QUITINASE Mo-chi1 [Moringa oleifera, LAM.], COM POLI-β-(1-4)-N-ACETIL-D-GLUCOSAMINE: UMA ABORDAGEM in silico. **DESAFIOS - Revista Interdisciplinar da Universidade Federal do Tocantins**, v. 5, n. 1, p. 111–120, 2018.

BLIN, K. et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. **Nucleic Acids Research**, v. 49, n. W1, p. W29–W35, 2 jul. 2021.

BROWN, J. R. Ancient horizontal gene transfer. **Nature Reviews Genetics**, v. 4, n. 2, p. 121–132, fev. 2003.

CARNEIRO, R.; COIMBRA, M. **Comparação entre Múltiplos Genomas para a Identificação de Sintenias**. [s.l.] Universidade de Brasília, 2010.

CHAE, H. Z. et al. Cloning and sequencing of thiol-specific antioxidant from mammalian brain: alkyl hydroperoxide reductase and thiol-specific antioxidant define a large family of antioxidant enzymes. **Proceedings of the National Academy of Sciences**, v. 91, n. 15, p. 7017–7021, 19 jul. 1994.

CHEN, Y.; LI, F.; WURTZEL, E. T. Isolation and Characterization of the Z-ISO Gene Encoding a Missing Component of Carotenoid Biosynthesis in Plants. **Plant Physiology**, v.

153, n. 1, p. 66–79, 3 maio 2010.

DOBSON, L.; REMÉNYI, I.; TUSNÁDY, G. E. CCTOP: a Consensus Constrained TOPology prediction web server. **Nucleic Acids Research**, v. 43, n. W1, p. W408–W412, 1 jul. 2015.

FERREIRA, L. S. DE S. et al. High-Quality Draft Genome Sequence of Pantanalinema sp. GBBB05, a Cyanobacterium From Cerrado Biome. **Frontiers in Ecology and Evolution**, v. 9, 2021.

GILCHRIST, C. L. M.; CHOOI, Y. H. Clinker & clustermap.js: Automatic generation of gene cluster comparison figures. **Bioinformatics**, v. 37, n. 16, p. 2473–2475, 2021.

GUO, F. et al. Adipocyte-derived PAMM suppresses macrophage inflammation by inhibiting MAPK signalling. **Biochemical Journal**, v. 472, n. 3, p. 309–318, 15 dez. 2015.

HALLGREN, J. et al. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. 2022.

KELLY, W. J. et al. The complete genome sequence of the rumen methanogen Methanobacterium formicicum BRM9. **Standards in Genomic Sciences**, v. 9, n. 1, p. 15, 8 dez. 2014.

KUMAR, S.; GADAGKAR, S. R. Efficiency of the Neighbor-Joining Method in Reconstructing Deep and Shallow Evolutionary Relationships in Large Phylogenies. **Journal of Molecular Evolution**, v. 51, n. 6, p. 544–553, 2 dez. 2000.

LEFORT, V.; DESPER, R.; GASCUEL, O. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program: Table 1. **Molecular Biology and Evolution**, v. 32, n. 10, p. 2798–2800, out. 2015.

LIBÓRIO, L.; RESENDE, V. H. Introdução aos bancos de dados biológicos. In: **BIOINFO - Revista Brasileira de Bioinformática e Biologia Computacional**. [s.l.] Alfahelix, 2021.

LIMA, F. P. **INFERENCIA BOOTSTRAP EM MODELOS DE REGRESSAO BETA**. [s.l.] Universidade Federal de Pernambuco, 2017.

MADIGAN, M. M. J. B. K. B. D. S. D. **Brock biology of microorganisms**. 14. ed. [s.l.] Pearson Education, Inc., 2016.

MARTIN, J. SAM (dependent) I AM: the S-adenosylmethionine-dependent methyltransferase fold. **Current Opinion in Structural Biology**, v. 12, n. 6, p. 783–793, 1 dez. 2002.

MATAFORA, V.; BACHI, A. Secret3D Workflow for Secretome Analysis. **STAR Protocols**, v. 1, n. 3, p. 100162, dez. 2020.

MEIER-KOLTHOFF, J. P.; GÖKER, M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. **Nature Communications**, v. 10, n. 1, p. 2182, 16 dez. 2019.

MOOSMANN, P. et al. Aranazoles: Extensively Chlorinated Nonribosomal Peptide–Polyketide Hybrids from the Cyanobacterium Fischerella sp. PCC 9339. **Organic Letters**, v. 20, n. 17, p. 5238–5241, 7 set. 2018.

MUHLSTEINOVA, R. et al. Polyphasic characterization of Trichocoleus desertorum sp. nov. (Pseudanabaenales, Cyanobacteria) from desert soils and phylogenetic placement of the genus Trichocoleus. **Phytotaxa**, v. 163, n. 5, p. 241, 31 mar. 2014.

NOBRE, C. A. S. **Isolamento, purificação e caracterização parcial da estrutura primária de uma ficobiliproteína da alga marinha vermelha Hypnea musciformus (WULFEN) LAMOUROUX**. [s.l.] Universidade Federal do Cerá, 2015.

OGAKI, M. B.; FURLANETO, M. C.; MAIA, L. F. Review: General aspects of bacteriocins. **Brazilian Journal of Food Technology**, v. 18, n. 4, p. 267–276, 2015.

PAUL, K. et al. The c-di-GMP Binding Protein YcgR Controls Flagellar Motor Direction and Speed to Affect Chemotaxis by a “Backstop Brake” Mechanism. **Molecular Cell**, v. 38, n. 1, p. 128–139, abr. 2010.

POLO, T. C. F.; MIOT, H. A. Aplicações da curva ROC em estudos clínicos e experimentais. **Jornal Vascular Brasileiro**, v. 19, p. 13–16, 2020.

PROSDÓCIMI, F.; MOREIRA, L. M. Genômica Comparativa. In: **Ciências genômicas: fundamentos e aplicações**. [s.l: s.n.]. p. 81–99.

RIBEIRO, D.; BRANCO, I.; CHOUPINA, A. B. Fatores moleculares no metabolismo fundamental de Phytophthora cinnamomi Molecular factors in the fundamental metabolism of Phytophthora cinnamomi. v. 44, p. 203–214, 2021.

SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. **Molecular Biology and Evolution**, jul. 1987.

SCHUBERT, H. L.; BLUMENTHAL, R. M.; CHENG, X. Many paths to methyltransfer: a chronicle of convergence. **Trends in Biochemical Sciences**, v. 28, n. 6, p. 329–335, jun.

2003.

SHANER, D. L. Acetohydroxyacid synthase inhibitors. **Reviews in Toxicology**, v. 1, n. 3–4, p. 69–110, 1997.

SILVA, F. F.; GONÇALVES, D. B.; LOPES, D. O. The use of bioinformatics tools to characterize a hypothetical protein from Penicillium rubens. **Genetics and Molecular Research**, v. 19, 2020.

SOUSA, W. B. **Estimadores de Máxima Verossimilhança : Casos que não satisfazem as condições de**. [s.l.] Universidade de Brasília- UnB, 2018.

SOUZA, L. DE N.; RHODEN, S. A.; PAMPHILE, J. A. A importância das ômicas como ferramentas para o estudo da prospecção de microrganismos: perspectivas e desafios. v. 2, p. 16–21, 2014.

SOUZA, D. **Uma Ferramenta Multiagente Baseada em Conhecimento para Anotação de Proteínas: um Estudo de Caso para o Fungo Saccharomyces cerevisiae**. [s.l.] Universidade de Brasília, 2014.

STOTHARD, P.; WISHART, D. S. Circular genome visualization and exploration using CGView. **Bioinformatics**, v. 21, n. 4, p. 537–539, 15 fev. 2005.

TAMURA, K. et al. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. **Molecular Biology and Evolution**, v. 28, n. 10, p. 2731–2739, 1 out. 2011.

TAMURA, K.; STECHER, G.; KUMAR, S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. **Molecular Biology and Evolution**, v. 38, n. 7, p. 3022–3027, 25 jun. 2021.

THAKUR, C. J. et al. Deciphering the functional role of hypothetical proteins from chloroflexus aurantiacs J-10-f1 using bioinformatics approach. **Molecular Biology Research Communications**, v. 9, n. 3, p. 129–139, 2020.

TRINDADE, V. M. T. et al. Avaliação Virtual Do Ponto Isoelétrico Da Caseína. **Dep. Bioquímica-ICBS-UFRGS**, p. 15, 2012.

VIJAYAKUMAR, S.; MENAKHA, M. Pharmaceutical applications of cyanobacteria—A review. **Journal of Acute Medicine**, v. 5, n. 1, p. 15–23, mar. 2015.

WYLIE, J. L.; WOROBEC, E. A. The OprB porin plays a central role in carbohydrate uptake

in Pseudomonas aeruginosa. **Journal of Bacteriology**, v. 177, n. 11, p. 3021–3026, jun. 1995.

ZHANG, B. et al. AAED1 modulates proliferation and glycolysis in gastric cancer. **Oncology Reports**, 7 jun. 2018.

ANEXO

Quadro 1. Resultados da previsão de domínios e famílias feitos pelas ferramentas SMART, INTERPRO, CATCH, PFAM E CDD de todas as HPs.

ID DA PROTEÍNA	SMART	InterPro	CATCH	Pfam	CD-Search
FWK01_00010	No domains	None predicted	No matches	DUF3107 Protein of unknown function/ HutP	PRK13683 superfamily
FWK01_00035	No domains	None predicted	No matches	No matches	No Domain
FWK01_00050	transmembrane region / Pfam:Peptidase_C39 / Pfam:PG_binding_1	protein binding/ ATP binding/ peptidase activity	4ry2A01/ 3k8uA01/ Haemolysin secretion ATP-binding protein	Peptidase_C39 Peptidase C39 family/ PG_binding_1 Putative peptidoglycan binding domain	Peptidase_C39_like super family/ PG_binding_1
FWK01_02825	low complexity	Signal peptide N-region	No matches	No matches	No Domain
FWK01_02830	Pfam:Phycobilisome	Phycobilisome, alpha/beta subunit	2vjtA00 (Phycocyanins)	Phycobilisome Phycobilisome protein	Globin-like super family- contain phycobilins
FWK01_02845	Pfam:NnrU / Pfam:PEMT	NnrU- 15 cis carotene	15-cis-zeta-carotene isomerase, chloroplastic / Protein-S-isoprenylcysteine O-methyltransferase	NnrU protein	Uncharacterized membrane protein
FWK01_02865	transmembrane region/ low complexity	None predicted	No matches	No matches	PRK13042 super family-superantigen-like protein SSL4; Reviewed;
FWK01_02890	Pfam:YlqD	YlqD protein	4dciG00	YlqD YlqD protein	YlqD protein
FWK01_02925	Pfam:HcyBio	HcyBio	No matches	HcyBio Homocysteine biosynthesis enzyme	HcyBio-Homocysteine biosynthesis enzyme, sulfur-incorporation

Quadro 1. Resultados da previsão de domínios e famílias feitos pelas ferramentas SMART, INTERPRO, CATCH, PFAM E CDD de todas as HPs.

ID DA PROTEÍNA	SMART	InterPro	CATCH	Pfam	CD-Search
FWK01_03355	Pfam:PNP_UDP_1	Nucleoside_phosphorylase_d / PNP_UDP_1	4wkbA00- Nucleoside phosphorylase domain/ 5-methylthioadenosine nucleosidase, S-adenosylhomocysteine nucleosidase/ Probable bifunctional mta/sah nucleosidase mtn	PNP_UDP_1- Phosphorylase superfamily	Pfs- Nucleoside phosphorylase [Nucleotide transport and metabolism];
FWK01_03360	low complexity	None predicted	No matches	DUF2417/ MitMem_reg-Maintenance of mitochondrial structure and function/ IL32-Interleukin 32	No Domain
FWK01_03365	coiled coil/ low complexity	coiled coil	No matches	HAMP HAMP domain/ Prefoldin Prefoldin subunit/ SpoOE-like Spo0E like sporulation regulatory protein/ HSBP1 Heat shock factor binding protein 1	No Domain
FWK01_03375	SCOP:d1qgua_	None predicted	No matches	Coq4 Coenzyme Q (ubiquinone) biosynthesis protein Coq4	No Domain
FWK01_03385	low complexity	None predicted	No matches	No matches	No Domain
FWK01_03410	transmembrane region	Carbohydrate-selective porin OprB	No matches	CD20 CD20-like family/ Imp-YgjV Bacterial inner membrane protein/ DUF2157 Predicted membrane protein (DUF2157)/ MHYT Bacterial signalling protein N terminal repeat	No Domain

Quadro 1. Resultados da previsão de domínios e famílias feitos pelas ferramentas SMART, INTERPRO, CATCH, PFAM E CDD de todas as HPs.

ID DA PROTEÍNA	SMART	InterPro	CATCH	Pfam	CD-Search
FWK01_03425	Carbohydrate-selective porin OprB	Carbohydrate-selective porin OprB	No matches	OprB Carbohydrate-selective porin, OprB family/ SLH S-layer homology domain	iron uptake porin
FWK01_04145	Pfam:SLH/ Pfam:OprB	None predicted	No matches	BBS1 Ciliary BBSome complex subunit 1	No Domain
FWK01_04195	SCOP domain	None predicted	no matches	DUF2007 Putative prokaryotic signal transducing protein	No Domain
FWK01_04225	low complexity/ transmembrane region	None predicted	no matches	no matches	gliding_GltJ super family
FWK01_06640	Pfam:YdjM	LexA-binding, inner membrane-associated putative hydrolase	no matches	YdjM LexA-binding, inner membrane-associated putative hydrolase	No Domain
FWK01_06650	No domains	None predicted	no matches	no matches	No Domain
FWK01_06655	Pfam:Imm1	Double-stranded DNA deaminase immunity protein DddI	no matches	Imm1 Immunity protein Imm1	No Domain
FWK01_06665	No domains	None predicted	no matches	TBPIP TBPIP/Hop2 winged helix domain	No Domain
FWK01_06680	SCOP domain	Sialidase_sf	no matches	BNR BNR/Asp-box repeat	No Domain

Quadro 1. Resultados da previsão de domínios e famílias feitos pelas ferramentas SMART, INTERPRO, CATCH, PFAM E CDD de todas as HPs.

ID DA PROTEÍNA	SMART	InterPro	CATCH	Pfam	CD-Search
FWK01_06690	SCOP domain	None predicted	no matches	DUF4259 Domain of unknown function	No Domain
FWK01_06695	No domains	None predicted	no matches	no matches	No Domain
FWK01_06700	low complexity/ coiled coil	Band_7/SPFH_dom_sf/ coiled coil	no matches	UPF0688 UPF0688 family	DUF5401 super family/ HflC super family
FWK01_06725	No domains	None predicted	no matches	no matches	No Domain
FWK01_06785	No domains	None predicted	no matches	Nudix_N_2 Nudix N-terminal/ DUF997 Protein of unknown function/ zf-C3HC4_2 Zinc finger, C3HC4 type (RING finger)	No Domain
FWK01_08650	Peptidase_C14	Tudor/ Chromo-like_dom_sf/	no matches	Domínio Peptidase_C14/ Agenet Domain	COG4249 super family-Uncharacterized protein, contains caspase domain
FWK01_08655	low complexity	None predicted	no matches	Família IDH/ Família FlaF	No Domain
FWK01_08665	HisKA Domain	sig_transdc_His_kin-like_/ His_kinase_dom/ HATPase_C/ GAF-like_dom_sf	Histidine kinase-like ATPase, C-terminal domain*	Ost4 Family/ Domínio GAF/ Domínio GvpK/ HisKA Domain / HATPase_c Domínio	BaeS- Signal transduction histidine kinase [Signal transduction mechanisms] / HK_sensor super family-Sensor domains of Histidine Kinase receptors
FWK01_08670	No domains	None predicted	no matches	DUF86 Família	No Domain
FWK01_08690	No domains	None predicted	no matches	no matches	No Domain
FWK01_08715	região transmembrana/ baixa complexidade	ARM-like	no matches	DUF5337 Família/ HEAT_2 Família	No Domain
FWK01_10130	no domains	SH3-related domain	no matches	Família Ntox33/ SH3_4 Domínio	No Domain

Quadro 1. Resultados da previsão de domínios e famílias feitos pelas ferramentas SMART, INTERPRO, CATCH, PFAM E CDD de todas as HPs.

ID DA PROTEÍNA	SMART	InterPro	CATCH	Pfam	CD-Search
FWK01_10145	no domains	None predicted	no matches	no matches	No Domain
FWK01_10150	no domains	None predicted	no matches	DUF1493 Domínio	No Domain
FWK01_10165	sequencia considerada muito curta	None predicted	no matches	Adenine_deam_C Family	metallo-dependent_hydrolases super family
FWK01_10220	região transmembrana/ SLH/ OprB	Carbohydrate-selective porin OprB/ SLH	no matches	Família SLH/ DHC_N1 Família/ Família OprB	por_somb super family-ron uptake porin
FWK01_10225	no domains	5peptide_repeat/	E3 ubiquitin-protein ligase SopA*	Repetição do pentapeptídeo/ Novirhabdo_Nv Família/	YjbI-Uncharacterized protein YjbI, contains pentapeptide repeats/ PRK15196 super family-type III secretion system effector PipB2
FWK01_10240	baixa complexidade/ Família zf-like	DUF6438	no matches	Família zf-like	No Domain
FWK01_10245	no domains	None predicted	no matches	no matches	No Domain
FWK01_10250	no domains	None predicted	no matches	Família Corona_NS2A/ DUF5596 Domínio/	No Domain
FWK01_10255	no domains	PROKAR_LIPOPROTEIN	no matches	no matches	No Domain
FWK01_10260	Pfam: MscS_porin	None predicted	Domínio de repetição de tetratricopeptídeo	MscS_porin Coiled-coil +90	COG1340 super family-Uncharacterized coiled-coil protein, contains DUF342 domain

Quadro 1. Resultados da previsão de domínios e famílias feitos pelas ferramentas SMART, INTERPRO, CATCH, PFAM E CDD de todas as HPs.

ID DA PROTEÍNA	SMART	InterPro	CATCH	Pfam	CD-Search
FWK01_10505	no domains	None predicted	no matches	no matches	No Domain
FWK01_10515	no domains	None predicted	no matches	no matches	No Domain
FWK01_10525	região transmembrana	None predicted	no matches	DUF3989 Família	No Domain
FWK01_10530	no domains	região transmembrana	no matches	DUF1691 Família / DUF2207 Família/ DUF3311 Família	No Domain
FWK01_15970	low complexity	Put_N_fixation	no matches	no matches	No Domain
FWK01_15980	transmembrane region	None predicted	no matches	no matches	No Domain
FWK01_15990	no domains	None predicted	no matches	UPF0688 Família	No Domain
FWK01_16010	PhyH Domain	None predicted	no matches	PhyH Domain	No Domain
FWK01_18075	no domains	None predicted	no matches	Domínio LnmK_N_HDF	No Domain
FWK01_18085	no domains	None predicted	no matches	FYVE_2 Família/ UPF0167 Família/ YgbA_NO Família	No Domain
FWK01_18090	Pfam: AhpC-TSA_2	Peroxiredoxin-like 2A/B/C	Selenoprotein U	AhpC-TSA_2 Domínio	Enzima antioxidante AhpC / TSA
FWK01_18165	low complexity/ SCOP:d1lkxa	Peroxiredoxin-like 2A/B/C	no matches	no matches	No Domain
FWK01_18185	no domains	None predicted	no matches	Peptidase_M91 Família	Superfamília M34_peptidase
FWK01_18190	no domains	None predicted	no matches	Transglut_prok Domain/ DUF771 Família / NUFIP1 Família	No Domain
FWK01_18215	low complexity/ coiled coil	None predicted	no matches	LMBR1 Família	No Domain

Quadro 1. Resultados da previsão de domínios e famílias feitos pelas ferramentas SMART, INTERPRO, CATCH, PFAM E CDD de todas as HPs.

ID DA PROTEÍNA	SMART	InterPro	CATCH	Pfam	CD-Search
FWK01_18230	baixa complexidade/ Pfam: DUF5122/ Pfam: VCBS/	None predicted	no matches	Família NAPRTase +	Super família NHL-Unidade de repetição de NHL de proteínas de hélice beta/VCBS-Repita o domínio em Vibrio, Colwellia, Bradyrhizobium e Shewanella
FWK01_18245	low complexity/ SCOP:d2ahja_	Integrin_alpha_N/ Delta_60_rpt	no matches	no matches	proteína da família da glicosil transferase;
FWK01_18270	low complexity / coiled coil	None predicted	no matches	Glyco_tran_28_C Domínio	Superfamília tipo Glicosiltransferase_GTB
FWK01_18275	Pfam: Glyco_tran_28_C	Glyco_tran_28_C	Monogalactosyldiacylglycerol synthase 3/ beta-glucosyltransferase	HTH_33 Domínio/ MGDG_synth Family/ Glyco_tran_28_C	Superfamília tipo Glicosiltransferase_GTB
FWK01_24935	transmembrane region/ coiled coil	Signal peptide H-region	Monogalactosyldiacylglycerol synthase 3/ Processive diacylglycerol beta-glucosyltransferase	BLOC1S3 Family/ LTXXQ Family / DUF4164 Família/ Família LuxE / DUF501 Família	No Domain
FWK01_26865	low complexity	Signal peptide H-region	no matches	no matches	No Domain
FWK01_26870	Pfam:JUPITER	None predicted	no matches	Família JUPITER	No Domain
FWK01_26875	transmembrane region/ low complexity	None predicted	no matches	no matches	No Domain
FWK01_26925	baixa complexidade/ WD40	WD40/YVTN_repeat-like_dom_sf	no matches	GUN4_N Domínio/ Família Nup160/ Ge1_WD40 / WD40 Repetir	Domínio WD40
FWK01_26930	WD40	quinoproteína amina desidrogenase / WD40	Tipo de repetição YVTN / quinoproteína amina desidrogenase	GUN4_N Domínio / WD40 Repetir	Domínio WD40

Quadro 1. Resultados da previsão de domínios e famílias feitos pelas ferramentas SMART, INTERPRO, CATCH, PFAM E CDD de todas as HPs.

ID DA PROTEÍNA	SMART	InterPro	CATCH	Pfam	CD-Search
FWK01_26940	no domains	LRR_dom_sf	Inibidor de ribonuclease	GUN4_N Domínio/ LRR_4 Repetição	Proteína da família STM4015
FWK01_26945	no domains	BETA-CITRYLGLUTAMATE SYNTHASE B/ Glutathione synthetase ATP-binding domain-like	no matches	Família ATPgrasp_ST	Proteína da família STM4014
FWK01_27080	Pfam: PG_binding_1	Peptidoglycan-bd-like	Superfamília semelhante a PGBD / PGBD	PG_binding_1 Domínio/ Glyco_hydro_19	PG_binding_1-Domínio de ligação de peptidoglicano putativo/ Super família tipo Lyz- domínios do tipo lisozima
FWK01_27085	SLH/ OprB	Porina seletiva de carboidratos OprB	no matches	SLH Family/ Spectrin Domain/ Família KfrA/ Família OprB// TolA_bind_tri	iron uptake porin
FWK01_27110	low complexity	none predicted	no matches	no matches	Super família PHA03247
FWK01_27140	no domains	none predicted	no matches	no matches	No Domain
FWK01_27145	no domains	none predicted	no matches	no matches	No Domain
FWK01_27155	no domains	HopA1 effector protein	no matches	Família Cerato-platanina/ HopA1 Domínio	Família de proteínas efetoras HopA1
FWK01_29045	Pfam: Ni_hydr_CYTB	Cyt_b561_bac/Ni-Hgenase	Redutase de tiorredoxina	Ni_hydr_CYTB Família/ Família PepSY_TM / Família Tmemb_55A	YdhU super family/ Thiosulfate reductase cytochrome b subunit

Quadro 2. Parâmetros físico-químicos para todas as HPs, utilizando o Expasy Protparam.

ID DA PROTEÍNA	ENTRADA UNIPROT	P.M	PI	CE	IE	IEC	IA	GRAVY
FWK01_00010	A0A098THZ1_9CYAN	9632.94	4.96	4470	30.20	stable	93.15	0. 017
FWK01_00035	K8GDT3_9CYAN	13989.94	4.38	10095	42.52	unstable	100.79	-0. 071
FWK01_00050	A0A2T1DX45_9CYAN	39351.04	9.54	56045	36.33	stable	112.82	0. 278
FWK01_02825	A0A2T1EGP8_9CYAN	11353.53	4.69	17990	43.08	unstable	64.56	-0. 270
FWK01_02830	K8GQ85_9CYAN	18171.00	5.14	13075	43.14	unstable	103.16	-0. 236
FWK01_02845	A0A2T1EGM5_9CYAN	26867.71	9.20	70930	29.07	stable	120.68	0. 571
FWK01_02865	G7XH68_ASPKW	9953.55	12.70	16500	78.78	unstable	83.09	0. 002
FWK01_02890	K8GJJ2_9CYAN	17168.69	4.95	5500	55.50	unstable	100.60	-0. 544
FWK01_02925	A0A2T1ENK8_9CYAN	42968.22	6 .02	48150	37.08	stable	96.98	-0. 040
FWK01_03355	A0A0V7ZHQ0_9CYAN	35817.97	5.56	33585	43.41	unstable	96.11	-0. 161
FWK01_03360	A0A3A1YQR1_9PAST	16806.71	8.74	23045	50.74	unstable	111.66	0. 361
FWK01_03365	A0A2T1DVU9_9CYAN	14912.74	9.64	5960	51.00	unstable	71.53	-0 .879
FWK01_03375	A0A2T1ENT8_9CYAN	19503.33	6,72	44015	42.55	unstable	99.70	-0. 121
FWK01_03385	A0A2T1EP36_9CYAN	36074.31	5.75	54110	46.03	unstable	101.31	-0. 1 59
FWK01_03410	A0A1Z4JAR3_LEPBY	11792.72	6.24	22460	32.33	stable	117.08	0. 537
FWK01_03425	A0A1Z4HNH5_9NOSO	57783.17	4.61	74050	22.42	stable	83.25	-0. 152
FWK01_04145	A0A4Q4X6D4_9PEZI	13405.21	5.40	1868	64.31	unstable	84.82	-0. 526
FWK01_04195	K9TJ71_9CYAN	7054.98	4.14	6990	57.32	unstable	106.46	-0 .034
FWK01_04225	A0A2K8STU6_9NOSO	36815.34	5.76	64065	57.43	unstable	80.87	-0. 510
FWK01_06640	F4XRR3_9CYAN	22844.34	9.79	36565	53.74	unstable	138.59	0. 690
FWK01_06650	W4M009_9BACT	24573.48	4.69	53985	39.51	stable	82.62	-0. 344

PM = peso moléculas; PI = potencial isoelétrico; CE = coeficiente de extinção; IE = índice de estabilidade; IEC = índice de estabilidade classificação; IA = índice alifático; GRAVY = índice de hidropaticidade média

Quadro 2. Parâmetros físico-químicos para todas as HPs, utilizando o Expasy Protparam.

ID DA PROTEÍNA	ENTRADA UNIPROT	P.M	PI	CE	IE	IEC	IA	GRAVY
FWK01_06655	I4D0E2_DESAJ	16983 .41	4..02	39670	57.29	unstable	66.33	-0. 463
FWK01_06665	A0A1Y6CJH3_9ALTE	17001.63	7.61	43680	55.89	unstable	94.79	-0. 313
FWK01_06680	A0A179D5C2_9BACT	12160.63	5.29	17085	53.71	unstable	62.34	-0. 571
FWK01_06690	B8HW45_CYAP4	9116.33	4.16	8480	75.37	unstable	101.25	-0. 300
FWK01_06695	A0A0C5WLW7_9GAMM	11545.28	6.83	1490	37.28	stable	108.22	0 .067
FWK01_06700	A0A251WHK9_9CYAN	53729.53	5.24	36565	49.25	unstable	94.95	-0. 557
FWK01_06725	Q7NNT6_GLOVI	14897.11	7.00	21095	40.82	unstable	90.38	-0. 053
FWK01_06785	A0A1Q8ZIA5_9CYAN	19020.14	9.82	57450	37.93	stable	77.72	-0. 326
FWK01_08650	K9X1Y9_9NOST	40024,31	4,17	102705	37,96	stable	37,96	-0, 400
FWK01_08655	K9WCN3_9CYAN	13.835,74	5,5	2980	39,34	stable	108,93	-0,002
FWK01_08665	A0A1E5QJN4_9CYAN	59025,67	5,3	53205	43,97	unstable	108,4	-0, 020
FWK01_08670	A0A2T1DSK9_9CYAN	24154,25	4,42	51450	54,95	unstable	90,62	-0,366
FWK01_08690	Q22Y61_TETTS	8549,52	5,51	10430	18,71	stable	59,19	-0, 607
FWK01_08715	A0A1Z4JKS2_LEPBY	24139,51	4,43	22460	44,35	unstable	109,1	0 ,025
FWK01_10130	A0A1E5QDL2_9CYAN	13329,21	9,61	20970	19,43	stable	90,92	-0, 384
FWK01_10145	A0A2T1E4P4_9CYAN	13617,56	5,37	8940	52,67	unstable	102,64	-0, 136
FWK01_10150	A0A2T1EG61_9CYAN	17373,68	4,87	36440	36,69	stable	90,56	-0, 322
FWK01_10165	A0A3S1AN37_CHLFR	3660,47	11,57	5500	26,79	stable	123,33	0, 153
FWK01_10220	A0A1Q8Z6K0_9CYAN	62789,3	4,81	60405	27,68	unstable	85,11	-0, 118

PM = peso moléculas; PI = potencial isoelétrico; CE = coeficiente de extinção; IE = índice de estabilidade; IEC = índice de estabilidade classificação; IA = índice alifático; GRAVY = índice de hidropaticidade média

Quadro 2. Parâmetros físico-químicos para todas as HPs, utilizando o Expasy Protparam.

ID DA PROTEÍNA	ENTRADA UNIPROT	P.M	PI	CE	IE	IEC	IA	GRAVY
FWK01_10225	A0A2T1CPW4_9CYAN	34398,95	6,92	14440	14,62	stable	111,22	0, 110
FWK01_10240	K9QAV5_9NOSO	20455,36	9,33	20065	46,85	unstable	79,34	-0, 388
FWK01_10245	A0A4Q5M8K5_9GAMM	8528,34	4,21	19480	34,42	stable	86,08	-0, 504
FWK01_10250	D9T4B9_MICAI	15126,21	4,62	13200	51,95	unstable	67,01	-0, 166
FWK01_10255	A0A1T5JFX7_9BACT	14115,97	5,27	52160	31,59	stable	70,96	-0 ,034
FWK01_10260	A0A6P7FD71_DIAVI	15618,06	6,25	28880	19,53	stable	82,56	-1, 005
FWK01_10505	A0A0K8J314_9FIRM	7117,24	9,18	8480	41,95	unstable	106,83	0 ,014
FWK01_10515	sem resultado	7723,21	4,7	1615	39,49	stable	117,39	0, 484
FWK01_10525	F4XXY5_9CYAN	7639,21	5,05	6990	25,41	stable	132,9	0 ,672
FWK01_10530	A0A2T1DM58_9CYAN	14246,83	6,13	27960	63,12	unstable	119,31	0, 656
FWK01_15970	A0A2T1E5M1_9CYAN	24894,81	4,26	34950	58,21	unstable	91,15	-0, 157
FWK01_15980	A0A110AXC0_9CYAN	6530,7	5,87	125	19,16	stable	118,62	0 ,911
FWK01_15990	A0A2T1DU87_9CYAN	29437,45	5,83	44585	45,83	unstable	85,45	-0,345
FWK01_16010	K9RGM5_9CYAN	30702,13	7,62	40715	43,56	unstable	88,31	-0 ,136
FWK01_18075	A0A1V4HQW9_9BACL	24888,19	6,17	7115	41,48	unstable	86,28	-0, 385
FWK01_18085	A0A1E5QR95_9CYAN	10005,55	6,25	7365	61,08	unstable	83,07	-0, 349
FWK01_18090	A0A2T1CEP9_9CYAN	29132,45	6,58	41160	40,49	unstable	94,66	-0 ,047
FWK01_18165	A0A401ILU5_APHSA	24226,41	5,02	40005	30,07	stable	96,05	-0, 298
FWK01_18185	A0A2T1ER12_9CYAN	26611,09	9,28	53860	43,39	unstable	61,22	-0, 669
FWK01_18190	A0A2T1DVK7_9CYAN	17684,14	6,08	53065	71,73	unstable	80, 00	-0, 676
FWK01_18215	S7MEA9_MYOBR	20079.91	9,28	40450	91,09	unstable	45,35	-2,024
FWK01_18230	A0A1Z4L483_NOSLI	74396,56	5,07	99725	12,1	stable	78,38	-0, 235
FWK01_18245	A0A2T1CDN6_9CYAN	9309,66	9,39	6990	50,23	unstable	105,93	-0 ,677

PM = peso moléculas; PI = potencial isoelétrico; CE = coeficiente de extinção; IE = índice de estabilidade; IEC = índice de estabilidade classificação; IA = índice alifático; GRAVY = índice de hidropaticidade média

Quadro 2. Parâmetros físico-químicos para todas as HPs, utilizando o Expasy Protparam.

ID DA PROTEÍNA	ENTRADA UNIPROT	P.M	PI	CE	IE	IEC	IA	GRAVY
FWK01_18270	A0A3S5K2F4_CHLFR	48632,2	8,06	59400	42,89	unstable	105,05	0, 004
FWK01_18275	A0A367QUI3_9NOSO	43938,5	6,42	61670	41,97	unstable	96,78	-0, 179
FWK01_24935	L8LVH1_9CYAN	16866,4	8,88	7115	63,86	unstable	104,87	-0, 288
FWK01_26865	K9VQR7_9CYAN	88822,7	5,51	70375	39,89	stable	90,41	-0,18
FWK01_26870	K9RKK8_9CYAN	12900,1	4,75	4,75	9,55	stable	63,38	0,261
FWK01_26875	K9RKK8_9CYAN	14656,6	9,7	4470	23,61	stable	83,79	0 ,283
FWK01_26925	K9VZF6_9CYAN	47443,8	5,97	81610	45,57	unstable	91,28	-0 ,275
FWK01_26930	K9VZF6_9CYAN	33885,4	6,11	47900	45,59	unstable	90,99	-0, 075
FWK01_26940	A0A0C1R0I6_9CYAN	47094,2	5,24	56880	39,69	stable	99,18	-0, 166
FWK01_26945	A0A0P4UI65_9CYAN	43617,8	8,48	69245	46,78	unstable	89,63	-0, 341
FWK01_27080	A0A1U7I3B5_9CYAN	30504,8	5,5	41035	30,17	stable	90,94	-0, 166
FWK01_27085	A0A2T1ER10_9CYAN	58257,6	4,59	67965	24,44	stable	82,36	-0, 166
FWK01_27110	K8GDQ1_9CYAN	52359	4,09	18450	56,07	unstable	77,05	-0, 314
FWK01_27140	A0A1Q4RRQ9_9CYAN	7932,69	3,98	4470	24,4	stable	93,42	-0 ,229
FWK01_27145	A0A1Q4RRQ9_9CYAN	8141,88	3,88	6990	45,21	unstable	90,81	-0 ,274
FWK01_27155	Q8YVP3_NOSS1	42010,4	6,05	50100	47,14	unstable	91,26	-0,34
FWK01_29045	B2IV05_NOSP7	12821,3	11,84	26470	38,59	stable	117,91	0,42

PM = peso moléculas; PI = potencial isoelétrico; CE = coeficiente de extinção; IE = índice de estabilidade; IEC = índice de estabilidade classificação; IA = índice alifático; GRAVY = índice de hidropaticidade média

Quadro 3. Determinação de localização subcelular de todas as HPs pelo Psortb, PSLpred e CELLO.

ID DA PROTEÍNA	Psortb	PSLpred	CELLO
FWK01_00010	Unknown	Inner-membrane Protein	Cytoplasmic
FWK01_00035	Cytoplasmic	Cytoplasmic Protein	Cytoplasmic
FWK01_00050	CytoplasmicMembrane	Inner-membrane Protein	InnerMembrane
FWK01_02825	Unknown	Extracellular Protein	Periplasmic
FWK01_02830	Cytoplasmic	Cytoplasmic Protein	Cytoplasmic
FWK01_02845	Inner-membrane Protein	CytoplasmicMembrane	InnerMembrane
FWK01_02865	CytoplasmicMembrane	Periplasmic Protein	Periplasmic
FWK01_02890	Cytoplasmic	Inner-membrane Protein	Cytoplasmic
FWK01_02925	Cytoplasmic	Cytoplasmic Protein	Cytoplasmic
FWK01_03355	Cytoplasmic	Cytoplasmic Protein	Cytoplasmic
FWK01_03360	CytoplasmicMembrane	Inner-membrane Protein	Cytoplasmic
FWK01_03365	Cytoplasmic	Cytoplasmic Protein	Cytoplasmic
FWK01_03375	Cytoplasmic	Inner-membrane Protein	Cytoplasmic
FWK01_03385	Cytoplasmic	Cytoplasmic	Cytoplasmic
FWK01_03410	Unknown	Extracellular Protein	Extracellular
FWK01_03425	Cellwall	Outer Membrane Protein	OuterMembrane
FWK01_04145	Cytoplasmic	Cytoplasmic Protein	Cytoplasmic
FWK01_04195	Unknown	Periplasmic Protein	Cytoplasmic
FWK01_04225	Cytoplasmic	Extracellular Protein	OuterMembrane
FWK01_06640	CytoplasmicMembrane	Inner-membrane Protein	InnerMembrane
FWK01_06650	CytoplasmicMembrane	Extracellular Protein	OuterMembrane/ Extracellular
FWK01_06655	Unknown	Outer Membrane Protein	Extracellular
FWK01_06665	Unknown	Cytoplasmic Protein	Cytoplasmic
FWK01_06680	Unknown	Extracellular Protein	Periplasmic
FWK01_06690	Cytoplasmic	Extracellular Protein	Cytoplasmic
FWK01_06695	Unknown	Cytoplasmic Protein	Cytoplasmic
FWK01_06700	Cytoplasmic	Cytoplasmic Protein	Cytoplasmic
FWK01_06725	CytoplasmicMembrane	Cytoplasmic Protein	Periplasmic
FWK01_06785	Cytoplasmic	Inner-membrane Protein	Cytoplasmic
FWK01_08650	Unknown	Cytoplasmic Protein	Cytoplasmic
FWK01_08655	CytoplasmicMembrane	Periplasmic Protein	Periplasmic
FWK01_08665	CytoplasmicMembrane	InnerMembrane	Cytoplasmic
FWK01_08670	Cytoplasmic	Cytoplasmic Protein	Cytoplasmic
FWK01_08690	Unknown	Outer Membrane Protein	Cytoplasmic/Periplasmic
FWK01_08715	Cytoplasmic	Cytoplasmic Protein	Cytoplasmic
FWK01_10130	Unknown	Periplasmic Protein	Membrana externa
FWK01_10145	Cytoplasmic	Cytoplasmic Protein	Cytoplasmic
FWK01_10150	Unknown	Inner-membrane Protein	Cytoplasmic

Quadro 3.Determinação de localização subcelular de todas as HPs pelo Psortb, PSLpred e CELLO.

ID DA PROTEÍNA	Psortb	PSLpred	CELLO
FWK01_10165	CytoplasmicMembrane	InnerMembrane	Cytoplasmic
FWK01_10220	Cellwall	InnerMembrane	Membrana externa
FWK01_10225	Unknown	Outer Membrane Protein	OuterMembrane
FWK01_10240	Unknown	Periplasmic Protein	Periplasmático
FWK01_10245	Unknown	Proteína Citoplasmática	Cytoplasmic
FWK01_10250	Cytoplasmic	Cytoplasmic Protein	Periplasmático
FWK01_10255	CytoplasmicMembrane	Outer Membrane Protein	Extracellular/ OuterMembrane/ Periplasmic
FWK01_10260	Unknown	Extracellular Protein	Extracellular
FWK01_10505	Cytoplasmic	Inner-membrane Protein	Cytoplasmic
FWK01_10515	Cytoplasmic	Inner-membrane Protein	Cytoplasmic
FWK01_10525	CytoplasmicMembrane	InnerMembrane	Citoplasmático/ Membrana Interna
FWK01_10530	CytoplasmicMembrane	Inner-membrane Protein	InnerMembrane/ Periplasmic
FWK01_15970	CytoplasmicMembrane	Periplasmic Protein	Cytoplasmic
FWK01_15980	CytoplasmicMembrane	Inner-membrane Protein	Cytoplasmic/ Periplasmic/ InnerMembrane
FWK01_15990	CytoplasmicMembrane	Periplasmic Protein	Periplasmic/ Cytoplasmic
FWK01_16010	Cytoplasmic	Cytoplasmic Protein	Cytoplasmic
FWK01_18075	Cytoplasmic	Cytoplasmic Protein	Cytoplasmic
FWK01_18085	Cytoplasmic	Cytoplasmic Protein	Cytoplasmic
FWK01_18090	CytoplasmicMembrane	Periplasmic Protein	Citoplasmático/ Periplasmic Protein/ CytoplasmicMembrane
FWK01_18165	Cytoplasmic	Cytoplasmic Protein	Cytoplasmic
FWK01_18185	Unknown	Periplasmic Protein	Periplasmático
FWK01_18190	Cytoplasmic	Inner-membrane Protein	Cytoplasmic
FWK01_18215	Unknown	Extracellular Protein	InnerMembrane/ Periplasmic
FWK01_18230	CytoplasmicMembrane	Proteína Extracelular	Extracelular
FWK01_18245	Cytoplasmic	InnerMembrane	Cytoplasmic
FWK01_18270	Cytoplasmic	Cytoplasmic Protein	Citoplasmático/ Membrana Interna
FWK01_18275	Unknown	Cytoplasmic Protein	Membrana Interna
FWK01_24935	Unknown	Periplasmic Protein	Periplasmic/ Membrana Interna/ Citoplasmático
FWK01_26865	OuterMembrane	Extracellular Protein	OuterMembrane
FWK01_26870	Unknown	Extracellular Protein	Extracellular
FWK01_26875	Unknown	Extracellular Protein	Extracellular
FWK01_26925	CytoplasmicMembrane	Cytoplasmic Protein	Membrana externa/ Citoplasmático/ Extracelular
FWK01_26930	Unknown	Periplasmic Protein	Periplasmático/ Membrana externa/ Extracelular

Quadro 3. Determinação de localização subcelular de todas as HPs pelo Psortb, PSLpred e CELLO.

ID DA PROTEÍNA	Psortb	PSLpred	CELLO
FWK01_26940	Cytoplasmic	Cytoplasmic Protein	Cytoplasmic
FWK01_26945	Cytoplasmic	Cytoplasmic Protein	Cytoplasmic
FWK01_27080	Extracellular	Cytoplasmic Protein	Cytoplasmic
FWK01_27085	Cellwall	Proteína Extracelular	Membrana externa
FWK01_27110	Unknown	Extracellular Protein	Periplasmic/ OuterMembrane/ Extracellular
FWK01_27140	Unknown	Extracellular Protein	Cytoplasmic
FWK01_27145	Unknown	Extracellular Protein	Cytoplasmic/ Extracellular
FWK01_27155	Unknown	Cytoplasmic Protein	Cytoplasmic
FWK01_29045	CytoplasmicMembrane	Inner-membrane Protein	InnerMembrane

Quadro 4. Predição de Hélices transmembrana *Deep*TMHMM e CCTOP, e caracterização de solubilidade pelo SOSUI para todas as HPs.

ID DA PROTEÍNA	SOSUI	CCTOP	TMHMM
FWK01_00010	SOLUBLE PROTEIN	1	0
FWK01_00035	SOLUBLE PROTEIN	1	1
FWK01_00050	MEMBRANE PROTEIN	6	4
FWK01_00850	SOLUBLE PROTEIN	1	0
FWK01_02825	SOLUBLE PROTEIN	1	0
FWK01_02830	SOLUBLE PROTEIN	1	0
FWK01_02845	MEMBRANE PROTEIN	4	5
FWK01_02865	MEMBRANE PROTEIN	1	1
FWK01_02890	SOLUBLE PROTEIN	0	0
FWK01_02925	SOLUBLE PROTEIN	2	0
FWK01_03355	SOLUBLE PROTEIN	3	0
FWK01_03360	MEMBRANE PROTEIN	1	0
FWK01_03365	SOLUBLE PROTEIN	1	0
FWK01_03375	SOLUBLE PROTEIN	1	0
FWK01_03385	MEMBRANE PROTEIN	3	0
FWK01_03410	MEMBRANE PROTEIN	8	3
FWK01_03425	SOLUBLE PROTEIN	1	0
FWK01_04145	SOLUBLE PROTEIN	1	0
FWK01_04195	SOLUBLE PROTEIN	1	0
FWK01_04225	MEMBRANE PROTEIN	5	1
FWK01_06640	MEMBRANE PROTEIN	5	4
FWK01_06650	SOLUBLE PROTEIN	1	0
FWK01_06655	SOLUBLE PROTEIN	1	0
FWK01_06665	SOLUBLE PROTEIN	1	0
FWK01_06680	SOLUBLE PROTEIN	1	0
FWK01_06690	SOLUBLE PROTEIN	1	0
FWK01_06695	SOLUBLE PROTEIN	1	0
FWK01_06700	SOLUBLE PROTEIN	2	0
FWK01_06725	SOLUBLE PROTEIN	1	0
FWK01_06785	MEMBRANE PROTEIN	1	0
FWK01_08650	SOLUBLE PROTEIN	1	0
FWK01_08655	SOLUBLE PROTEIN	1	0
FWK01_08665	MEMBRANE PROTEIN	1	1
FWK01_08670	SOLUBLE PROTEIN	1	1
FWK01_08690	SOLUBLE PROTEIN	1	0
FWK01_08715	MEMBRANE PROTEIN	1	1
FWK01_09050	SOLUBLE PROTEIN	1	1
FWK01_10130	MEMBRANE PROTEIN	0	0
FWK01_10145	SOLUBLE PROTEIN	0	0
FWK01_10150	SOLUBLE PROTEIN	2	0

Quadro 4. Predição de Hélices transmembrana *Deep*TMHMM e CCTOP, e caracterização de solubilidade pelo SOSUI.

ID DA PROTEÍNA	SOSUI	CCTOP	TMHMM
FWK01_10165	SOLUBLE PROTEIN	1	0
FWK01_10220	SOLUBLE PROTEIN	1	1
FWK01_10225	SOLUBLE PROTEIN	2	0
FWK01_10240	SOLUBLE PROTEIN	1	0
FWK01_10245	SOLUBLE PROTEIN	2	0
FWK01_10250	SOLUBLE PROTEIN	1	0
FWK01_10255	MEMBRANE PROTEIN	1	0
FWK01_10260	SOLUBLE PROTEIN	1	0
FWK01_10505	SOLUBLE PROTEIN	1	0
FWK01_10515	SOLUBLE PROTEIN	1	0
FWK01_10525	MEMBRANE PROTEIN	1	1
FWK01_10530	MEMBRANE PROTEIN	3	3
FWK01_15970	SOLUBLE PROTEIN	3	0
FWK01_15980	MEMBRANE PROTEIN	2	2
FWK01_15990	SOLUBLE PROTEIN	1	0
FWK01_16010	SOLUBLE PROTEIN	1	0
FWK01_18075	SOLUBLE PROTEIN	1	0
FWK01_18085	SOLUBLE PROTEIN	0	0
FWK01_18090	SOLUBLE PROTEIN	0	0
FWK01_18165	SOLUBLE PROTEIN	1	0
FWK01_18185	SOLUBLE PROTEIN	2	0
FWK01_18190	SOLUBLE PROTEIN	1	0
FWK01_18215	MEMBRANE PROTEIN	1	0
FWK01_18230	SOLUBLE PROTEIN	18	0
FWK01_18245	SOLUBLE PROTEIN	1	0
FWK01_18270	SOLUBLE PROTEIN	1	0
FWK01_18275	SOLUBLE PROTEIN	1	0
FWK01_20895	SOLUBLE PROTEIN	1	1
FWK01_20920	MEMBRANE PROTEIN	1	1
FWK01_20940	SOLUBLE PROTEIN	3	0
FWK01_20960	SOLUBLE PROTEIN	1	0
FWK01_20985	SOLUBLE PROTEIN	2	0
FWK01_24935	MEMBRANE PROTEIN	1	1
FWK01_26865	SOLUBLE PROTEIN	11	0
FWK01_26870	SOLUBLE PROTEIN	1	0
FWK01_26875	MEMBRANE PROTEIN	1	1
FWK01_26925	SOLUBLE PROTEIN	1	0
FWK01_26930	SOLUBLE PROTEIN	1	0
FWK01_26940	SOLUBLE PROTEIN	1	0
FWK01_26945	SOLUBLE PROTEIN	1	0
FWK01_27080	SOLUBLE PROTEIN	1	0

FWK01_27085	SOLUBLE PROTEIN	6	0
FWK01_27110	SOLUBLE PROTEIN	0	0
FWK01_27140	SOLUBLE PROTEIN	0	1
FWK01_27145	SOLUBLE PROTEIN	0	1
FWK01_27155	SOLUBLE PROTEIN	0	0
FWK01_27435	SOLUBLE PROTEIN	1	0
FWK01_29045	MEMBRANE PROTEIN	2	2