



UNIVERSIDADE DO ESTADO DA BAHIA
DEPARTAMENTO DE CIÊNCIAS EXATAS E DA TERRA
CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

EVELYN SOUZA FERREIRA

PLASTICOME: UM MÉTODO COMPUTACIONAL PARA MINERAÇÃO
EM GENOMAS FÚNGICOS A FIM DE ENCONTRAR POTENCIAIS
ENZIMAS QUE DEGRADAM PLÁSTICOS

SALVADOR

2024

EVELYN SOUZA FERREIRA

PLASTICOME: UM MÉTODO COMPUTACIONAL PARA MINERAÇÃO EM
GENOMAS FÚNGICOS A FIM DE ENCONTRAR POTENCIAIS ENZIMAS QUE
DEGRADAM PLÁSTICOS

Monografia apresentada ao curso de Sistemas de Informação do Departamento de Ciências Exatas e da Terra da Universidade do Estado da Bahia - UNEB, como requisito à obtenção do grau de Bacharel em Sistemas de Informação. Área de Concentração: Bioinformática

Orientador: Alexandre Rafael Lenz

SALVADOR

2024

EVELYN SOUZA FERREIRA

PLASTICOME: UM MÉTODO COMPUTACIONAL PARA MINERAÇÃO EM
GENOMAS FÚNGICOS A FIM DE ENCONTRAR POTENCIAIS ENZIMAS QUE
DEGRADAM PLÁSTICOS

Monografia apresentada ao curso de Sistemas de Informação do Departamento de Ciências
Exatas e da Terra da Universidade do Estado da Bahia - UNEB, como requisito à
obtenção do grau de Bacharel em Sistemas de Informação. Área de Concentração:
Bioinformática

Aprovada em:

BANCA EXAMINADORA

Alexandre Rafael Lenz (Orientador)
Universidade do Estado da Bahia – UNEB

Membro da Banca 1
IES do Membro da Banca 1

Membro da Banca 2
IES do Membro da Banca 2

Dedico este trabalho ao meu noivo e melhor amigo, Pedro Cândido. A você, que esteve ao meu lado durante toda essa jornada, mesmo nos momentos em que me senti exausta e desmotivada. O seu apoio constante e a sua crença em minha capacidade foram essenciais para que eu superasse os desafios e encontrasse forças para continuar.

AGRADECIMENTOS

Gostaria de expressar meus sinceros agradecimentos ao meu orientador, Alexandre Rafael Lenz, por seu apoio, orientação e dedicação ao longo deste trabalho. Também quero estender meus agradecimentos ao professor Diego Gervasio Frías Suárez, cujas correções e contribuições significativas foram essenciais para melhorar este trabalho. Agradeço, igualmente, aos meus colegas de classe, cujo apoio e colaboração foram fundamentais nesta jornada da graduação. Suas contribuições foram inestimáveis e tornaram essa experiência acadêmica mais leve e ainda mais enriquecedora.

"A vida entrelaçada é uma jornada brilhante e cativante para as vidas ocultas dos fungos - os grandes conectores do mundo vivo - e seus papéis surpreendentes e íntimos na vida humana, com o poder de curar nossos corpos, expandir nossas mentes e nos ajudar a enfrentar nossos problemas ambientais mais urgentes"

(Merlin Sheldrake)

RESUMO

No cenário da crescente preocupação com o acúmulo de plásticos, a micorremediação emerge como uma solução promissora, dado o papel crucial desempenhado pelos fungos na reciclagem de matéria orgânica. A chave para a micorremediação reside na identificação dos fungos mais adequados para combater poluentes específicos. Essa seleção pode ocorrer empiricamente, expondo os fungos aos poluentes e observando os resultados ao longo do tempo, ou através de abordagens *in silico* que exploram os genomas dos fungos para identificar seu potencial de degradar poluentes. Levando em consideração que algumas vias metabólicas podem permanecer inativas nos fungos por várias gerações, manifestando-se apenas quando expostas a substâncias específicas, descobrir esses genes "adormecidos" em experimentos de bancada pode demandar considerável tempo. No entanto, o Plasticome identifica essa potencialidade ao final da análise das proteínas anotadas do genoma, o que leva em torno de 3 horas.

O Plasticome compreende um banco de dados e um pipeline capaz de identificar enzimas com potencial para degradar plásticos em genomas fúngicos. Primeiramente, foi construído um banco de dados baseado em uma revisão da literatura, contendo enzimas fúngicas comprovadamente eficazes na degradação de diferentes tipos de plásticos, com destaque para o Polietileno (PE) e o Poliestireno (PS), os plásticos mais produzidos e utilizados globalmente. O PE, em particular, demonstrou ter o maior potencial de degradação por fungos, sendo consumido por cinco tipos de enzimas: peroxidases de manganês, peroxidases versáteis, lacases, peroxidases de lignina e cutinases.

O pipeline permite a seleção de um genoma a partir do *GenBank*, seguido pelo download de todas as sequências de proteínas do fungo escolhido. Ele compreende três etapas: 1) identificação de Enzimas Ativas em Carboidratos (*CAZy*) usando o software dbCan (v4.0.0). Os resultados do dbCan são filtrados para manter apenas as CAZymes associadas às atividades de degradação de plásticos encontradas no banco de dados; 2) os resultados filtrados da etapa 1 são submetidos ao ECPred (v1.1) para identificar os números EC (*Enzyme Commission*), que classificam enzimas com base nas reações químicas que catalisam; e 3) em seguida, o *BLAST* identifica a similaridade com as enzimas já conhecidas do banco de dados.

Finalmente, os resultados são consolidados em gráficos que relacionam as enzimas mineradas e os respectivos tipos de plásticos que elas poderiam degradar. Por exemplo, foi observado

que o *Aspergillus brasiliensis* IFM 66951 possui potencial para degradação de PE, com três lacases EC 1.10.3.2 da família AA1 e duas cutinases EC 3.1.1.74 da família CE5. Nossa abordagem *in silico* economiza tempo e custo em comparação com testes empíricos, contribuindo para o avanço urgente que a micorremediação necessita. O Plasticome está disponível no repositório do G2BC no *GitHub*.

Palavras-chave: Bioinformática. Micoinformática. Degradação de plásticos. Biorremediação. Fungos. Micorremediação;

ABSTRACT

In the scenario of growing concern about plastic accumulation, mycoremediation emerges as a promising solution, given the crucial role played by fungi in recycling organic matter. The key to mycoremediation lies in identifying the most suitable fungi to combat specific pollutants. This selection can occur empirically, exposing the fungi to pollutants and observing the results over time, or through *in silico* approaches that explore the genomes of fungi to identify their potential to degrade pollutants. Considering that some metabolic pathways may remain inactive in fungi for several generations, manifesting only when exposed to specific substances, discovering these “dormant” genes in bench experiments can take considerable time. However, Plasticome identifies this potential at the end of the analysis of the annotated proteins of the genome, which takes around 3 hours.

Plasticome comprises a database and a pipeline capable of identifying enzymes with potential to degrade plastics in fungal genomes. Firstly, a database was built based on a literature review, containing fungal enzymes proven to be effective in degrading different types of plastics, with emphasis on Polyethylene (PE) and Polystyrene (PS), the most produced and used plastics globally. PE, in particular, has shown to have the highest potential for degradation by fungi, being consumed by five types of enzymes: manganese peroxidases, versatile peroxidases, laccases, lignin peroxidases, and cutinases.

The pipeline allows the selection of a genome from *GenBank*, followed by the download of all protein sequences of the chosen fungus. It comprises three steps: 1) identification of Carbohydrate-Active Enzymes (CAZy) using the dbCan software (v4.0.0). The dbCan results are filtered to keep only the CAZymes associated with the activities of plastic degradation found in the database; 2) the filtered results from step 1 are submitted to ECPred (v1.1) to identify the EC numbers (*Enzyme Commission*), which classify enzymes based on the chemical reactions they catalyze; and 3) then, *BLAST* identifies the similarity with the already known enzymes from the database.

Finally, the results are consolidated into graphs that relate the mined enzymes and the respective types of plastics they could degrade. For example, it was observed that *Aspergillus brasiliensis* IFM 66951 has potential for PE degradation, with three laccases EC 1.10.3.2 from the AA1 family and two cutinases EC 3.1.1.74 from the CE5 family. Our *in silico* approach saves time and cost compared to empirical tests, contributing to the

urgent advancement that mycoremediation needs. Plasticome is available in the G2BC repository on *GitHub*.

Keywords: Bioinformatics. Mycoinformatics. Plastic degradation. Bioremediation. Fungi. Mycoremediation;

LISTA DE FIGURAS

Figura 1 – Filhote de albatroz morto com plástico no estômago	18
Figura 2 – Estrutura dos fungos	19
Figura 3 – Esquema do dbCAN-sub para permitir a anotação de substrato no nível de subfamília de CAZyme	27
Figura 4 – Método de predição usado pelo ECPred	29
Figura 5 – Alinhamento de duas proteínas usando BLAST	30
Figura 6 – Modelo do banco de dados de enzimas	39
Figura 7 – Representação gráfica para os passos da pipeline definida	41
Figura 8 – Representação de três proteínas em um arquivo fasta	41
Figura 9 – Exemplo de saída do dbcan, em um arquivo separado por tabulação . .	42
Figura 10 – Exemplo de saída após filtro de CAZyS	43
Figura 11 – Exemplo de saída da fase do ECPred, em um arquivo separado por tabulação	43
Figura 12 – Exemplo de saída do filtro de números EC, em um arquivo separado por tabulação	44
Figura 13 – Exemplo de saída do alinhamento com Blast	45
Figura 14 – Gráfico enviado ao usuário relacionando enzimas e tipos de plástico . .	45
Figura 15 – Interface da ferramenta	47
Figura 16 – Arquitetura do sistema	48
Figura 17 – Consulta 1: Resultado gráfico com potenciais enzimas encontradas no <i>Neurospora tetrasperma</i> FGSC 2508.	52
Figura 18 – Consulta 2: Resultado gráfico com potenciais enzimas encontradas no <i>Aspergillus brasiliensis</i> IFM 66951.	53
Figura 19 – Consulta 3: Resultado gráfico com potenciais enzimas encontradas no <i>Fusarium oxysporum</i> Fo47.	54
Figura 20 – Consulta 4: Resultado gráfico com potenciais enzimas encontradas no <i>Ganoderma sinense</i> ZZ0214-1.	55
Figura 21 – Consulta 5: Resultado gráfico com potenciais enzimas encontradas no <i>Pleurotus pulmonarius</i> PM_ss5.	57

SUMÁRIO

1	INTRODUÇÃO	14
2	REFERENCIAL TEÓRICO	17
2.1	Referencial teórico da área de aplicação	17
2.1.1	Poluição Ambiental por Acúmulo de Plásticos na Natureza	17
2.1.2	Microrremediação	18
2.1.3	Genômica de Fungos	21
2.1.4	Enzimas (EC number e CAZymes)	21
2.1.5	Enzimas com potencial para degradação de plásticos	23
2.1.6	Contribuições do trabalho na área de aplicação	23
2.2	Referencial teórico da área computacional	24
2.2.1	Bioinformática	24
2.2.1.1	Pipeline de Bioinformática	24
2.2.1.2	Mineração em Genomas	25
2.2.1.3	Repositórios de dados e ferramentas utilizadas	25
2.2.1.3.1	<i>GenBank</i>	25
2.2.1.3.2	<i>Ferrementa dbCAN</i>	26
2.2.1.3.3	<i>Ferramenta ECPred</i>	27
2.2.1.3.4	<i>BLAST</i>	28
2.2.2	Ferramentas para construção do pipeline	30
2.2.2.1	Linguagem Python e biblioteca Biopython	31
2.2.2.2	RabbitMQ	31
2.2.2.3	Celery	32
2.2.2.4	Flask	32
2.2.2.5	PostgreSQL	33
2.3	Trabalhos correlatos	33
3	METODOLOGIAS	35
3.1	Metodologia de pesquisa	35
3.2	Metodologia de desenvolvimento	36

4	RESULTADOS	38
4.1	Banco de Dados de Enzimas	38
4.1.1	Modelagem do banco de dados	38
4.1.2	Enzimas coletadas para o banco de dados	39
4.2	Plasticome: Método Computacional	40
4.3	Aplicação Web	46
4.3.1	Arquitetura da Aplicação	47
4.3.2	Verificação de funcionalidade	49
4.3.2.1	Resultados da verificação	51
5	TRABALHOS FUTUROS	58
6	CONSIDERAÇÕES FINAIS	59
	REFERÊNCIAS	60

1 INTRODUÇÃO

A bioinformática desempenha um papel fundamental na análise e interpretação de dados genômicos, oferecendo métodos e ferramentas computacionais para extrair informações relevantes de sequências genéticas. No contexto da crescente preocupação com a poluição causada pelo acúmulo de plásticos no meio ambiente (MARTIN et al., 2020), surgem novos desafios e oportunidades para a aplicação da bioinformática na identificação de fungos com capacidade de degradação de plásticos.

Os fungos são organismos eucariotos que desempenham papéis essenciais nos ecossistemas, incluindo a reciclagem de materiais orgânicos (YADAV et al., 2021). Dentre os fungos, há aqueles capazes de degradar os plásticos, oferecendo uma alternativa promissora para a mitigação desse problema ambiental. No entanto, a identificação das proteínas envolvidas nesse processo de degradação em diferentes espécies de fungos requer abordagens avançadas, devido à complexidade e ao tamanho dos genomas.

Nesse contexto, surge a necessidade de desenvolver métodos computacionais que possam analisar e minerar informações nos genomas de fungos, a fim de identificar as proteínas relacionadas à degradação de plásticos. Esses métodos podem se basear em técnicas de alinhamento de sequências (MIYAUCHI et al., 2020), que permitem comparar sequências de proteínas conhecidas com sequências genômicas desconhecidas, a fim de identificar semelhanças e inferir a presença de proteínas de interesse, também pode ser feito o uso ou adaptação de *softwares* e bancos de dados preexistentes com a finalidade de identificar as enzimas que degradam plásticos.

A utilização de métodos computacionais nessa análise proporciona uma abordagem eficiente e escalável, uma vez que os genomas de fungos podem conter um grande número de genes e informações que seriam inviáveis de serem analisadas manualmente. Para compreender a magnitude desses dados, é importante considerar a unidade de medida utilizada para descrever o comprimento de moléculas de ácidos nucleicos, como o DNA, que é o kilobase (kbp). Um kilobase representa 1.000 pares de bases de DNA, sendo que cada par é composto por duas bases químicas ligadas entre si que podem ser: adenina (A),

citossina (C), guanina (G) ou timina (T). Essas bases se unem de forma específica: A com T e C com G, formando um "degrau da escada do DNA", um genoma fúngico na média possui 42.300 Megabases (Mbp)(MOHANTA; BAE, 2015).

Essa vasta quantidade de informações genômicas demanda o uso de abordagens computacionais para a análise e interpretação dos dados. Os métodos computacionais permitem explorar a diversidade genômica dos fungos de forma mais eficiente, identificando genes e proteínas com potencial de degradação de plásticos. Além disso, esses métodos oferecem uma oportunidade de descobrir novas proteínas que desempenham um papel importante nesse processo.

Nesta pesquisa, busca-se explorar a aplicação de um método computacional baseado em alinhamento de sequências para a mineração em genomas de fungos, visando identificar as proteínas relacionadas à degradação de plástico. São utilizadas técnicas de bioinformática e análise genômica para investigar os genomas de diferentes espécies de fungos em busca dessas proteínas de interesse.

Diante disso, surgem as seguintes perguntas norteadoras:

- Quais proteínas estão presentes nos fungos e têm capacidade de degradar plásticos?
- Existem métodos computacionais específicos para identificar essas proteínas no genoma de um fungo?
- Quais são esses métodos e como funcionam?

O objetivo geral deste trabalho é desenvolver um método computacional capaz de minerar genomas de fungos para encontrar enzimas com potencial de degradação de plásticos. Para alcançar esse objetivo geral, são estabelecidos os seguintes objetivos específicos:

- Realizar uma revisão da literatura existente sobre o tema.
- Propor um método computacional para a identificação dos genes alvo nos genomas de fungos.
- Construir um banco de dados de enzimas com ação comprovada na degradação de plásticos
- Validar o método proposto utilizando o banco de dados proposto no objetivo anterior.

Os resultados esperados para esta pesquisa são:

- Revisão sistemática da literatura.
- Aplicação web implementando o método computacional proposto.
- Banco de dados de enzimas com atividade comprovada na degradação de plásticos.
- Validação qualitativa da aplicação web.

Ao responder a essas perguntas e atingir esses objetivos, espera-se contribuir para o avanço do conhecimento científico na área de biodegradação de plásticos por fungos.

2 REFERENCIAL TEÓRICO

O presente capítulo sustenta a fundamentação e o embasamento conceitual desta monografia. Por meio de uma revisão bibliográfica criteriosa, busca-se explorar os principais estudos, teorias e conceitos que constituem a base teórica do tema em questão.

2.1 REFERENCIAL TEÓRICO DA ÁREA DE APLICAÇÃO

Nesta seção, será apresentado o referencial teórico que sustenta o desenvolvimento deste trabalho, com foco nos conceitos fundamentais das sequências genômicas fúngicas e na aplicação de pipelines na bioinformática. As referências utilizadas foram obtidas por meio de uma extensa busca em bases de dados científicas reconhecidas na área da bioinformática, bem como por recomendação de especialistas atuantes nesse campo.

2.1.1 Poluição Ambiental por Acúmulo de Plásticos na Natureza

O acúmulo de plásticos na natureza é uma preocupação cada vez mais presente em todo o mundo. O descarte inadequado e a falta de reciclagem desse material têm contribuído para um aumento significativo de sua presença nos ecossistemas terrestres e aquáticos. Esse acúmulo de plásticos tem impactos negativos na fauna, flora e também na saúde humana, afetando a biodiversidade, a qualidade da água e do solo.

O crescente e contínuo acúmulo de plásticos provenientes de ações e práticas humanas em ecossistemas aquáticos causa interrupções diretas e indiretas na estrutura e nas funções desses ecossistemas, comprometendo seus serviços e valores. Além de causar a morte de animais marinhos por ingestão ou enredamento, o plástico tem a capacidade de contaminar toda a cadeia alimentar, uma vez que sua degradação é lenta e persistente.(THUSHARI; SENEVIRATHNA, 2020)

Quando o fotógrafo americano Chris Jordan visitou o Atol de Midway, no meio do Oceano Pacífico, em setembro de 2009, sua intenção era documentar os alarmantes níveis de poluição nos oceanos. Após capturar fotografias de enormes montanhas de lixo, Jordan estava em busca de uma abordagem diferente para ilustrar a magnitude do

problema do consumo excessivo de plástico, quando soube de uma ilha remota, localizada a 2.100 km a noroeste de Honolulu, no Havaí, coberta por milhares de aves mortas com seus estômagos repletos de resíduos plásticos do cotidiano, como tampas de garrafas e escovas de dentes (figura 1) (BBC, 2023).

Figura 1 – Filhote de albatroz morto com plástico no estômago



Fonte: Chris Jordan

Para enfrentar esse problema global, diversas soluções estão em discussão, incluindo o desenvolvimento de plásticos biodegradáveis (FLURY; NARAYAN, 2021). Além disso, a micorremediação surge como uma abordagem promissora, embora atualmente careça de meios suficientes para ser aplicada em escala. A falta de ferramentas para coletar informações e orientar a micorremediação na área de degradação de plásticos é um dos principais obstáculos a serem superados nessa busca por soluções eficazes.

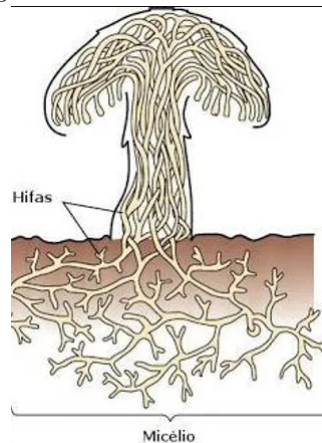
2.1.2 Micorremediação

Os fungos são seres vivos com atividade metabólica que possuem habilidades notáveis para explorar, investigar e decompor substâncias resistentes. Eles utilizam combinações de enzimas e substâncias ácidas poderosas para desintegrar componentes como lignina (presente na madeira), rochas, petróleo bruto, plásticos de poliuretano e até mesmo o explosivo TNT. Algumas espécies de fungos encontradas em locais contaminados por resíduos nucleares, como Chernobyl, são extremamente resistentes à radiação e podem até mesmo utilizar a radiação como fonte de energia, assim como as plantas utilizam a luz solar. Em situações de destruição, como após uma explosão nuclear, cogumelos como o matsutake são os primeiros organismos vivos a emergir (SHELDRAKE, 2020).

Essa forma de usar fungos para devolver um ambiente contaminado por poluentes a um estado menos poluído se chama **micorremediação**, o termo refere-se ao uso de micélios de fungos na biorremediação. Um dos papéis primários dos fungos num ecossistema é a decomposição, que é efetuada pelo micélio. O micélio segrega enzimas e ácidos extracelulares que decompõem a lignina e a celulose, os principais constituintes das fibras vegetais. Trata-se de compostos orgânicos constituídos por cadeias longas de carbono e hidrogênio, estruturalmente semelhantes a muitos poluentes orgânicos.

Micélio é o corpo dos fungos, que consiste de uma massa de ramificação formada por um conjunto de hifas emaranhadas (figura 2). O micélio é responsável pela absorção de nutrientes do ambiente em que se encontra. Ele faz isso em um processo de duas etapas, primeiro as hifas secretam enzimas no alimento ou sobre ele, quebrando polímeros biológicos em unidades menores, como monômeros. Esses monômeros são então absorvidos pelo micélio por difusão facilitada e transporte ativo. O micélio desempenha um papel vital nos ecossistemas terrestres e aquáticos, pois contribui para a decomposição do material vegetal e seu crescimento libera dióxido de carbono de volta para a atmosfera.

Figura 2 – Estrutura dos fungos



Fonte: Blog só biologia

Os fungos possuem uma ampla diversidade de apetites, mas existem certos materiais que eles não conseguem decompor, a menos que sejam forçados a fazê-lo. Eles não produzem enzimas desnecessárias. Algumas enzimas ou caminhos metabólicos podem permanecer inativos nos genomas dos fungos por várias gerações e se manifestarem apenas quando expostos a uma situação de necessidade (SHELDRAKE, 2020). Um exemplo disso é o treinamento do fungo *Pleurotus* para digerir bitucas de cigarro, que são um dos itens mais

comumente descartados no mundo. Essas bitucas contêm resíduos tóxicos que dificultam sua decomposição natural. No entanto, ao longo do tempo, o fungo *Pleurotus* (um cogumelo comestível conhecido popularmente como Hiratake) foi gradualmente alimentado com bitucas de cigarro usadas, eliminando outras opções de alimento. Com o tempo, o fungo "aprendeu" a utilizar exclusivamente as bitucas como fonte de alimento, e seu micélio começou a crescer e se alimentar desses resíduos.

Quando o fungo *Pleurotus* digere as bitucas de cigarro, pode ser necessário ativar uma enzima que não era usada há muito tempo ou utilizar uma nova enzima para essa função específica. Muitas enzimas fúngicas não são específicas para um único composto, o que significa que uma única enzima pode ser útil para metabolizar diferentes compostos com estruturas semelhantes. Isso é especialmente vantajoso quando se trata de poluentes tóxicos, como os encontrados nas bitucas de cigarro, que se assemelham aos subprodutos da decomposição da lignina. Portanto, oferecer bitucas de cigarro ao fungo *Pleurotus* representa um desafio comum para ele. A micologia radical baseia-se na capacidade dos fungos em degradar uma ampla variedade de compostos. No entanto, nem sempre é possível prever quais substâncias uma determinada cepa fúngica será capaz de metabolizar. Por exemplo, algumas cepas de fungos podem evitar ou crescer através das substâncias consumindo-as enquanto outras podem parar de crescer próximo a elas. Isso demonstra a diversidade de chaves enzimáticas presentes nos fungos: algumas prontas para uso imediato e outras ocultas dentro de seus genomas necessitando uma análise mais detalhada.

A chave da micorremediação reside em identificar a espécie de fungo adequada para combater um poluente específico. Essa seleção pode ser realizada empiricamente, de forma a envolver a coleta de várias espécies de fungos, a exposição a diferentes poluentes e a observação dos resultados ao longo do tempo. Esse método, embora valioso, pode ser demorado, uma vez que requer experimentação e coleta de dados a longo prazo para determinar a eficácia de uma espécie de fungo em particular para um determinado poluente. Alternativamente, e de forma mais eficiente, a seleção pode ser realizada por meio de abordagens *in silico*. Essas abordagens envolvem análises computacionais usando a genômica dos fungos para identificar genes ou características que sugerem a capacidade de degradar um poluente específico.

2.1.3 Genômica de Fungos

A genômica de fungos é um campo de estudo dedicado à análise e compreensão dos genomas de diferentes espécies de fungos. O genoma de um fungo é composto por todo o seu conjunto de DNA, que contém todas as informações genéticas necessárias para o funcionamento e desenvolvimento do organismo.

Este DNA é transcrito em RNA através de um processo chamado transcrição. O RNA, por sua vez, é traduzido em proteínas, conforme o dogma central da biologia molecular. Estas proteínas podem ser enzimas, proteínas transportadoras, entre outras, que desempenham funções vitais para o organismo, este processo é inerente e crucial para qualquer ser vivo (CRICK, 1970).

A genômica de fungos envolve o uso de técnicas avançadas de sequenciamento de DNA, análise bioinformática e estudos funcionais para explorar e decifrar as informações genéticas presentes nos genomas fúngicos. Essa abordagem permite identificar e estudar genes específicos relacionados a características importantes dos fungos, como a capacidade de decompor materiais orgânicos, produzir metabólitos úteis ou causar doenças em plantas e animais.

O estudo genômico dos fungos fornece *insights* valiosos sobre a diversidade genética, evolução, adaptação e interações dos fungos com o ambiente, o que é de extrema importância para avaliar o potencial biotecnológico das espécies e compreender como elas funcionam e podem ser utilizadas de forma benéfica (MORGAN, 2021).

Em particular para a aplicação da micorremediação da poluição com plásticos é necessário encontrar no genoma dos fungos as características das enzimas que tem a capacidade de realizar a degradação de plásticos.

2.1.4 Enzimas (EC number e CAZymes)

As enzimas CAZy (*Carbohydrate-Active EnZymes*) são uma classificação de enzimas que estão envolvidas no metabolismo de carboidratos. Elas desempenham um papel fundamental na quebra, síntese e modificação de diferentes tipos de carboidratos, como celulose, quitina, amido e outros polissacarídeos.

A classificação CAZy foi desenvolvida pelo grupo de Glicogenômica no AFMB (*Laboratoire Architecture et Fonction des Macromolécules Biologiques*) em Marselha, França. com o objetivo de categorizar e nomear as enzimas com atividade em carboidratos. Essa classificação é baseada na sequência de aminoácidos e na estrutura das enzimas, bem como na função que desempenham no metabolismo dos carboidratos.

As enzimas CAZy são divididas em várias classes e cada classe possui famílias com um código numérico identificador, como pode ser visto na tabela 2, cada uma das classes é associada a um grupo específico de enzimas com funções semelhantes. Essas classes incluem: *Glycoside Hydrolases (GHs)*, *GlycosylTransferases (GTs)*, *Carbohydrate Esterases (CEs)* e outras (CAZY, 2023).

Os números EC (*Enzyme Commission*) são um sistema de classificação numérica para enzimas, baseado nas reações químicas que catalisam. Cada enzima é descrita por um número de quatro dígitos, onde o primeiro dígito representa a classe da enzima e os outros três fornecem informações adicionais sobre a reação catalisada.

Tabela 0 – As 5 classes de CAZymes e seus Números EC

CAZyme	Número EC
Auxiliary Activities (AAs)	1.-.-
Carbohydrate Esterases (CEs)	3.1.1.-
Glycoside Hydrolases (GHs)	3.2.1.-
Glycosyl Transferases (GTs)	2.4.-.-
Polysaccharide Lyases (PLs)	4.2.2.-

Fonte: O autor

Essas enzimas são encontradas em uma variedade de organismos, incluindo bactérias, fungos, plantas e animais. Elas desempenham papéis essenciais em processos biológicos importantes, como a digestão de alimentos, a degradação de biomassa vegetal, a biossíntese de carboidratos estruturais e a modificação de carboidratos em células.

A classificação CAZy é uma metodologia valiosa para os pesquisadores que estudam enzimas envolvidas no metabolismo de carboidratos. Ela ajuda na compreensão da diversidade enzimática, na identificação de novas enzimas e na investigação de suas funções e aplicações biotecnológicas. O estudo e a compreensão das enzimas CAZy têm implicações importantes em áreas como bioenergia, produção de bioplásticos, produção de enzimas industriais e desenvolvimento de terapias farmacológicas.

2.1.5 Enzimas com potencial para degradação de plásticos

Os fungos tem potencial para degradar diferentes tipos de plásticos, incluindo plásticos à base de petróleo e plásticos biodegradáveis. Vários estudos foram realizados para investigar a degradação de plásticos por fungos, e os resultados mostram que diferentes espécies de fungos têm a capacidade de degradar diferentes tipos de plásticos. Por exemplo, o fungo *Penicillium simplicissimum* foi capaz de degradar HDPE (polietileno de alta densidade) com a ajuda de um tratamento prévio de irradiação UV. Além disso, o fungo *Fusarium solani* foi capaz de degradar plásticos PU (poliuretano) em compostagem a 25°C. Existem várias atividades enzimáticas necessárias para a degradação de diferentes tipos de plásticos, principalmente hidrolíticas, mas também oxidoredutases. É importante destacar que as enzimas relevantes para a degradação de plásticos incluem lacases (AA1), peroxidases de classe II (AA2) e esterases (CE5). Além disso, as atividades de proteases/peptidases (EC 3.4.-) também podem ser aplicáveis à degradação de plásticos (DALY et al., 2021).

Durante a realização desta pesquisa, foram identificados os seguintes plásticos que apresentaram algum grau de degradação causada por fungos:

Tabela 1 – Plásticos encontrados com algum grau de degradação fúngica.

Nome do plástico	Sigla
Polietileno	PE
Polietileno tereftalato	PET
polietileno de alta densidade	HDPE
Policloreto de Vinila	PVC
Poliuretano	PU/PUR

Fonte: O autor

2.1.6 Contribuições do trabalho na área de aplicação

Atualmente, a área da micorremediação carece de métodos eficazes ou repositórios abrangentes para armazenar informações relacionadas à degradação de plásticos por meio de atividade enzimática em fungos. A contribuição essencial desse projeto consiste na criação de um repositório que unifica e arquiva esses dados, fornecendo um valioso recurso para a comunidade científica. Além disso, por meio deste projeto, também são geradas novas informações e conhecimentos a respeito do processo de degradação de plásticos por fungos, enriquecendo, assim, a base de dados disponível.

2.2 REFERENCIAL TEÓRICO DA ÁREA COMPUTACIONAL

Nesta seção, será apresentado o referencial teórico da área computacional, fornecendo a base necessária para o desenvolvimento do método proposto para a mineração de genomas. Serão explorados os métodos e técnicas utilizados por outros estudos na área, bem como os programas e ferramentas computacionais que possuem funcionalidades relevantes para a análise e mineração de genomas. Ao explorar este referencial teórico da área computacional, busca-se obter um amplo conhecimento sobre as abordagens, técnicas e ferramentas que são utilizadas no campo da bioinformática para a mineração de genomas. Com base nessa compreensão, será possível desenvolver um método computacional robusto e eficiente, aproveitando as melhores práticas e recursos disponíveis para alcançar os objetivos propostos neste trabalho.

2.2.1 Bioinformática

A Bioinformática é uma disciplina que combina a biologia com a tecnologia da informação para resolver problemas relacionados à análise de dados biológicos, e para que isso seja possível alguns conceitos são extremamente importantes.

2.2.1.1 Pipeline de Bioinformática

o conceito de pipeline na bioinformática descreve uma abordagem sistemática e automatizada para processar e analisar grandes volumes de dados genômicos. Essa abordagem envolve a criação de uma sequência de etapas interconectadas, que podem incluir pré-processamento dos dados, alinhamento de sequências, anotação funcional, identificação de variantes genéticas, entre outras tarefas. A utilização de pipelines na bioinformática agiliza o fluxo de trabalho, aumenta a reprodutibilidade dos resultados e permite a integração de diferentes ferramentas e recursos para a análise genômica (BINI et al., 2021).

O objetivo de se ter um pipeline é organizar e analisar informações genéticas complexas utilizando uma combinação de ferramentas de informática, matemática e estatística, para que essas informações possam ser interpretadas por profissionais da área biológica. Um dos passos mais importantes em um pipeline de bioinformática é a mineração de genomas, para extrair informações do genoma dos organismos alvo.

2.2.1.2 Mineração em Genomas

Mineração em Genomas é o processo de extrair informações úteis e conhecimento a partir de grandes conjuntos de dados genômicos. Essa técnica é amplamente utilizada na bioinformática para identificar padrões e relações em dados genômicos, com o objetivo de entender melhor a biologia molecular e celular. Uma aplicação importante da mineração em genomas é a análise de dados de vírus. Um estudo realizado pela Universidade Federal de Ciências da Saúde de Porto Alegre utilizou técnicas de mineração de dados para analisar uma base pública de dados de genoma do vírus influenza (CORRÊA, 2017). Os resultados mostraram que a mineração de dados é uma ferramenta promissora para a descoberta de novos conhecimentos na área da saúde.

Em resumo, a mineração em genomas é uma técnica poderosa que permite aos pesquisadores extrair informações valiosas a partir de grandes conjuntos de dados genômicos. Essa técnica tem diversas aplicações na bioinformática e na pesquisa médica, e está sendo cada vez mais utilizada para desenvolver novas terapias e tratamentos. Além da mineração em genomas para extrair e analisar informações também são usados softwares e repositórios de dados biológicos. Na próxima seção serão descritos os softwares e repositórios utilizados no pipeline desse trabalho.

2.2.1.3 Repositórios de dados e ferramentas utilizadas

Os bancos de dados e softwares consolidados desempenham um papel fundamental na área da bioinformática, fornecendo acesso a informações valiosas e ferramentas essenciais para análise e interpretação de dados biológicos. Neste tópico, exploraremos algumas dessas ferramentas que serão utilizadas no desenvolvimento dessa pesquisa.

2.2.1.3.1 *GenBank*

O NCBI (*National Center for Biotechnology Information*) é uma instituição responsável por fornecer acesso e ferramentas para explorar e analisar informações biológicas e genéticas. Ele hospeda o GenBank, além de outros bancos de dados relacionados à genômica, proteômica, estruturas tridimensionais e literatura científica.

O NCBI e o GenBank são recursos essenciais para a comunidade científica na

área da genética e biologia molecular. O GenBank é um banco de dados público e global que armazena sequências genéticas, incluindo sequências de DNA e RNA, de diversas espécies. É considerado um dos maiores repositórios de dados genéticos do mundo.

O GenBank possui um papel fundamental na divulgação e compartilhamento de informações genéticas, permitindo que pesquisadores e cientistas de todo o mundo tenham acesso a sequências genéticas de diferentes organismos. Isso possibilita estudos comparativos, análises evolutivas, descobertas de novas espécies e investigações sobre as bases genéticas de doenças, entre outras aplicações.

No contexto dos fungos, o GenBank disponibiliza um vasto número de genomas completos de fungos para a comunidade científica. De acordo com a pesquisa realizada, atualmente, existem 14.697 (consulta em 14/06/2023) genomas completos de fungos disponíveis nesse banco de dados. Essas informações genéticas são valiosas para estudos sobre a diversidade fúngica, biotecnologia, evolução e outros aspectos relacionados aos fungos.

2.2.1.3.2 *Ferrementa dbCAN*

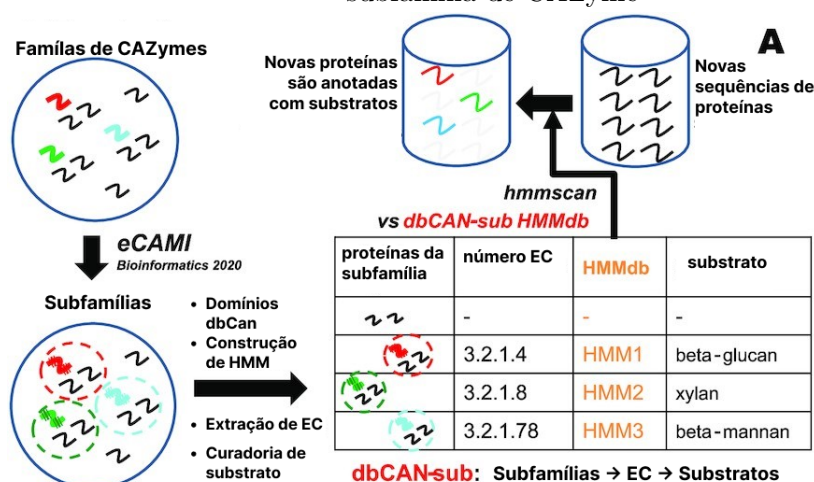
O dbCAN (ZHENG et al., 2023) é um servidor da web para anotação automatizada de enzimas ativas em carboidratos (CAZymes) em genomas. Ele usa ferramentas de bioinformática para identificar CAZymes e prever seus substratos. O processo de anotação começa com a entrada de sequências de proteínas em formato FASTA no servidor da web do dbCAN. Em seguida, o dbCAN executa uma análise de domínio de CAZymes usando o banco de dados de modelos ocultos de Markov (HMMdb) para identificar CAZymes em sequências de proteínas.

O dbCAN3, a versão mais recente, combina várias ferramentas para melhorar a anotação de CAZymes e inclui o CGC-Finder, uma ferramenta para identificar clusters de genes de CAZymes (CGCs) em genomas. Ele também possui um banco de dados de modelos ocultos de Markov (HMMdb) para prever substratos de CAZymes em níveis de subfamília e níveis de CGC. O dbCAN3 usa uma abordagem de votação para prever substratos de CGCs, considerando todas as CAZymes que fazem parte de um CGC e que tiveram seus substratos previstos pelo dbCAN-sub. Cada previsão de substrato é considerada como um “voto”, e o substrato que receber a maioria dos “votos” é então

previsto como o substrato do CGC.

O dbCAN é amplamente utilizado em várias aplicações, como bioenergia, microbioma, nutrição, agricultura e reciclagem global de carbono. Ele é capaz de identificar CAZymes em sequências de proteínas de diferentes organismos, incluindo bactérias, fungos, plantas e animais. O dbCAN também é capaz de prever substratos de CAZymes, o que é útil para entender a função biológica dessas enzimas em diferentes contextos. O dbCAN é uma ferramenta importante para a análise de genomas e metagenomas em várias áreas de pesquisa, incluindo biotecnologia, microbiologia, ecologia e biologia molecular.

Figura 3 – Esquema do dbCAN-sub para permitir a anotação de substrato no nível de subfamília de CAZyme



Fonte: (ZHENG et al., 2023)

2.2.1.3.3 Ferramenta ECPred

O ECPred é uma poderosa ferramenta de bioinformática voltada para a previsão dos códigos EC. Além disso, prevê as funções das proteínas que não foram ainda caracterizadas, uma tarefa de suma importância na bioinformática. Apesar de muitos métodos e ferramentas terem sido desenvolvidos para a classificação de enzimas, a maioria deles se concentra em classes funcionais específicas e níveis hierárquicos específicos do sistema de classificação EC (DALKIRAN et al., 2018).

No ECPred, cada número EC é tratado como uma classe distinta, o que resulta em um modelo de aprendizado independente para cada uma. Além disso, o ECPred incorpora uma abordagem de previsão hierárquica que explora a estrutura hierárquica da nomenclatura EC, que se assemelha a uma árvore. Para a construção do conjunto de

treinamento negativo do mesmo número EC, são usadas proteínas que não têm nenhuma anotação de função enzimática, ou seja, não são enzimas, e aquelas anotadas com outros números EC, ou seja, proteínas pertencentes a diferentes famílias enzimáticas.

O ECPred possui uma arquitetura abrangente que fornece previsões para um total de 858 números EC. Isso inclui 6 classes principais, 55 subclasses, 163 sub-subclasses e 634 classes de substratos. A ferramenta ECPred foi desenvolvida em Java e está disponível de duas formas: como uma ferramenta autônoma com seu código-fonte acessível no GitHub (<https://github.com/cansyl/ECPred>) e como um serviço web acessível através do site (<https://ecpred.kansil.org/>). Essa ferramenta oferece previsões de função enzimática com base em probabilidades para sequências de proteínas não caracterizadas em todos os cinco níveis da classificação EC. É uma contribuição significativa no campo da bioinformática, permitindo uma análise detalhada e previsões confiáveis de funções enzimáticas a partir de sequências de proteínas previamente não caracterizadas. (DALKIRAN et al., 2018)

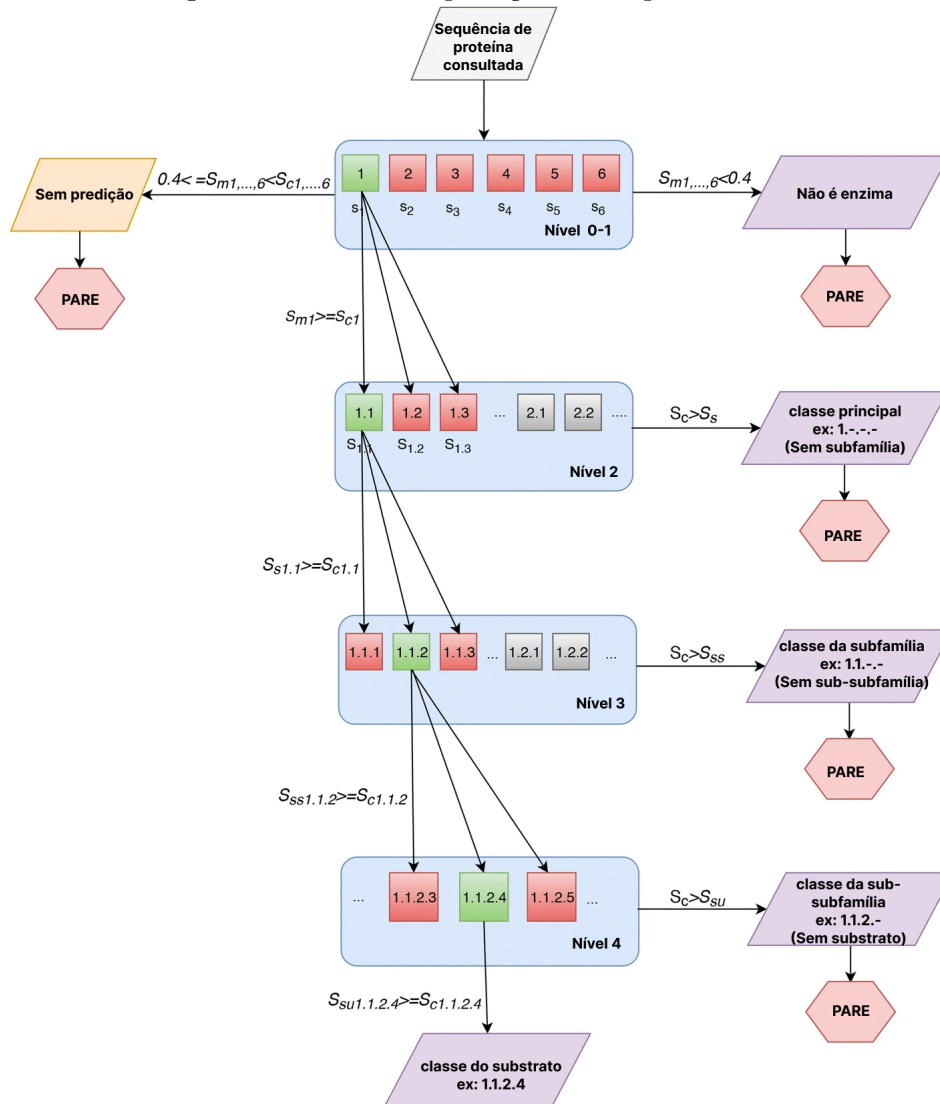
2.2.1.3.4 BLAST

O BLAST (*Basic Local Alignment Search Tool*) é uma ferramenta amplamente utilizada na área de bioinformática para realizar alinhamentos de sequências de DNA, RNA e proteínas. Ele desempenha um papel fundamental na análise comparativa de sequências biológicas, permitindo identificar similaridades e relações evolutivas entre diferentes organismos.

O BLAST utiliza algoritmos eficientes para comparar uma sequência de interesse com um banco de dados contendo diversas sequências já conhecidas. O objetivo é encontrar regiões de alta similaridade entre a consulta e as sequências do banco, indicando possíveis relações funcionais, estruturais ou evolutivas. A principal utilidade do BLAST está na anotação funcional de sequências desconhecidas, permitindo inferir funções biológicas a partir de similaridades com sequências já caracterizadas. Além disso, o BLAST auxilia na identificação de homólogos em diferentes espécies, possibilitando estudos comparativos entre organismos.

A ferramenta BLAST oferece diferentes tipos de busca, como BLASTp (comparação de sequências de proteínas), BLASTn (comparação de sequências de DNA), BLASTx (tradução de sequências de DNA para proteínas e posterior comparação) e muitos outros.

Figura 4 – Método de predição usado pelo ECPred



Fonte: (DALKIRAN et al., 2018)

Essas variantes do BLAST são adaptadas para atender às necessidades específicas dos pesquisadores e facilitar a análise de diferentes tipos de dados biológicos (MCGINNIS; MADDEN, 2004).

Na Figura 5, pode-se observar o resultado de um alinhamento realizado com a proteína *Pho4p* de *S. cerevisiae* (levedura). A sequência de consulta possui 184 aminoácidos, enquanto a sequência de referência contém 269 aminoácidos. O alinhamento apresenta um *bit-score* de 83.6, um *E-value* de 3e-14 e uma porcentagem de identidade de 44% (61/136). Além disso, a porcentagem de similaridade é de 53% (73/136) e a porcentagem de lacunas é de 13% (18/136). Isso indica que, além dos 44% de resíduos idênticos, existem mais 9% de resíduos que são similares, e 13% de lacunas no alinhamento. Esses dados sugerem uma significativa similaridade entre as duas sequências. Para melhor compreensão desses

dados, segue uma breve explicação dos termos utilizados:

- *Bit-score*: Representa a qualidade do alinhamento. Quanto maior o *bit-score*, melhor é o alinhamento.
- *E-value*: Refere-se à expectativa estatística de encontrar um alinhamento igualmente bom ou melhor por acaso. Um E-value menor indica que é menos provável que o alinhamento seja devido ao acaso, tornando o alinhamento mais significativo.
- Identidade (%): Corresponde à porcentagem de resíduos (aminoácidos ou nucleotídeos) que são idênticos no alinhamento.
- Similaridade(%): É a porcentagem de resíduos (aminoácidos ou nucleotídeos) que são similares no alinhamento. Isso inclui resíduos idênticos e também resíduos que não são idênticos, mas possuem propriedades químicas semelhantes. Por exemplo, a leucina e a isoleucina são consideradas similares porque ambas são aminoácidos hidrofóbicos.
- Lacunas (%): Representa a porcentagem de lacunas no alinhamento. As lacunas são introduzidas para otimizar o alinhamento quando um resíduo em uma sequência não tem correspondente na outra sequência. Uma alta porcentagem de lacunas pode indicar que as duas sequências têm um nível baixo de similaridade.

Figura 5 – Alinhamento de duas proteínas usando BLAST

Alinhamento BLAST - Pho4p (*S. cerevisiae*)

Bit-Score = 83.6 bits (205), E-value = 3e-14, Método: ajuste de matriz composicional

Identidade = 61/136 (44%), Similaridade = 73/136 (53%), Lacunas = 18/136 (13%)

Consulta	184	KPKPKQYPKVLPSNSTRRI	SPVTAKTSSSAEGVVVASESP	VIAPHSRSLSKRRSS	243
		KP P P+ ILPSN+ +R P S V+ AS+SPVI P+ + RS			
Referência	269	KPAPG-LPRFILPSNNPQRQLPP	SDS-----VIHASQSPVIKPN	YAGKPPGFVSAR	322

Fonte: O Autor

2.2.2 Ferramentas para construção do pipeline

A área de bioinformática tem se beneficiado cada vez mais do desenvolvimento de aplicações web para análise e processamento de dados biológicos. Nesse contexto, a combinação da linguagem de programação Python com a biblioteca Biopython tem

se destacado como uma poderosa ferramenta para manipulação e análise de sequências biológicas. Além disso, o uso de sistemas de mensageria como o RabbitMQ e *frameworks* como o Celery têm facilitado o gerenciamento e distribuição de tarefas computacionalmente intensivas. Integrado a essas tecnologias, o micro *framework* Flask oferece uma sólida estrutura para o desenvolvimento de aplicativos web direcionados à bioinformática. Nesta seção, exploraremos com mais detalhes as ferramentas mais relevantes para este trabalho.

2.2.2.1 Liguagem Python e biblioteca Biopython

Python é uma linguagem de programação de alto nível, amplamente utilizada em ambientes comerciais e acadêmicos. Sua sintaxe é fácil de aprender, possui recursos de programação orientada a objetos e uma ampla variedade de bibliotecas. Python pode ser usado para interface com código otimizado escrito em C e C++ o que a torna ótima escolha para programação científica (PYTHON, 2023).

Biopython é uma biblioteca que possui um conjunto de ferramentas gratuitas para computação biológica escritas em Python por uma equipe internacional de desenvolvedores. É um esforço colaborativo distribuído para desenvolver bibliotecas e aplicativos Python que atendam às necessidades do trabalho atual e futuro em bioinformática. O código-fonte está disponível sob a licença Biopython, que é extremamente liberal e compatível com quase todas as licenças do mundo (BIOPYTHON, 2023).

Na área de bioinformática, Python e Biopython são importantes porque oferecem uma ampla gama de ferramentas e recursos para análise de sequências, estruturas de proteínas, consulta a bancos de dados biológicos, genética populacional, filogenia e muito mais. Isso permite realizar análises complexas e avançadas de maneira eficiente e eficaz (COCK et al., 2009).

2.2.2.2 RabbitMQ

RabbitMQ é um software de corretagem de mensagens de código aberto que originalmente implementou o protocolo avançado de enfileiramento de mensagens (AMQP). É usado como um intermediário para lidar com o processamento de mensagens em segundo plano e comunicação entre serviços. Ele aceita e encaminha mensagens, permitindo que os aplicativos se conectem a ele para transferir uma mensagem ou mensagens. Uma mensagem

pode incluir qualquer tipo de informação.

RabbitMQ é amplamente utilizado, leve e fácil de implantar no ambiente local e na nuvem. Ele suporta vários protocolos de mensagens e pode ser implantado em configurações distribuídas e federadas para atender aos requisitos de alta escala e alta disponibilidade. Por esse motivo foi escolhido para lidar com o processamento assíncrono das tarefas e a comunicação entre os serviços (RABBITMQ, 2023).

2.2.2.3 Celery

O Celery é um sistema de fila de tarefas distribuído e assíncrono em código aberto, geralmente usado em desenvolvimento de software para tratar tarefas em segundo plano. Ele é particularmente útil em situações em que é necessário processar tarefas demoradas ou intensivas em recursos de forma eficiente, sem bloquear o aplicativo principal. Ele utiliza uma arquitetura cliente-servidor, onde um ou mais servidores executam as tarefas e o cliente envia as solicitações de execução.

Essa biblioteca é altamente escalável e flexível, permitindo que as tarefas sejam distribuídas em vários nós de processamento, o que torna o Celery ideal para lidar com cargas de trabalho intensivas e que exigem alto desempenho. Além disso, ele oferece recursos avançados, como agendamento de tarefas, monitoramento e gerenciamento de filas, o que facilita a implementação de sistemas robustos e eficientes (CELLERY, 2023).

2.2.2.4 Flask

Flask é um micro *framework* que utiliza a linguagem Python para criar aplicativos Web de forma simples e eficiente (FLASK, 2023). Ele é ideal para quem busca simplicidade, rapidez, soluções para projetos pequenos e aplicações robustas. Alguns dos benefícios do flask no desenvolvimento de aplicações:

- Simplicidade: Por possuir apenas o necessário para o desenvolvimento de uma aplicação, um projeto escrito com Flask é mais simples se comparado aos *frameworks* maiores.
- Flexibilidade: O Flask é conhecido por sua flexibilidade. Ele oferece uma abordagem minimalista para o desenvolvimento web em Python, fornecendo apenas as

funcionalidades essenciais para criar aplicações web.

- **Extensibilidade:** O Flask possui uma vasta comunidade de desenvolvedores que contribuem com extensões e pacotes que podem ser facilmente integrados às aplicações.
- **Performance:** Por ser um *microframework*, o Flask tem uma performance superior quando comparado a outros *frameworks* mais pesados.
- **Compatibilidade:** O Flask é compatível com diversas bibliotecas e ferramentas utilizadas em bioinformática, facilitando a integração de diferentes componentes.

2.2.2.5 PostgreSQL

O PostgreSQL é um sistema de gerenciamento de banco de dados relacional de código aberto, altamente confiável e poderoso. Ele é amplamente utilizado em diferentes domínios, incluindo bioinformática, devido às suas características e capacidades avançadas.

Na área de bioinformática, onde grandes quantidades de dados genômicos, proteômicos e outros dados biológicos são gerados, é crucial ter um sistema robusto e eficiente para armazenar, consultar e analisar essas informações. O PostgreSQL se destaca como uma escolha popular nesse contexto devido às suas principais vantagens.

Além disso, o PostgreSQL oferece recursos avançados, como suporte a consultas complexas e capacidade de extensão. Outra vantagem do PostgreSQL é a sua escalabilidade. À medida que a quantidade de dados aumenta, é essencial ter um sistema que possa crescer e lidar com essa demanda sem comprometer o desempenho (WISESO et al., 2020).

2.3 TRABALHOS CORRELATOS

- **ANTISMASH** (GALAL et al., 2023): Foi desenvolvido em Python e também usa o Redis como banco de dados e o Werkzeug e o Jinja2 como ferramentas auxiliares. O ANTISMASH é um software de análise de genomas que identifica e caracteriza biossínteses de moléculas secundárias em genomas bacterianos e fúngicos. Ele é um software de código aberto disponível sob a licença GNU (acessível em: <https://github.com/antismash/antismash>)
- **2ndFind** (ZHANG et al., 2021): É uma ferramenta de análise de sequências genômicas bacterianas baseada na comparação de sequências desconhecidas com um banco

de dados contendo elementos conhecidos, como transposons e integrons. A ferramenta utiliza algoritmos de busca de similaridade para encontrar regiões homólogas (que possuem uma origem ancestral em comum), permitindo a identificação de possíveis elementos genéticos móveis. É uma ferramenta WEB que pode ser acessada em (<https://biosyn.nih.go.jp/2ndfind/>), porém não possui o código aberto e não fornece informações detalhadas sobre o desenvolvimento.

- **TGFAM-FINDER** (ZHOU et al., 2021): Uma ferramenta de análise de genomas foi desenvolvida para identificar novas famílias de genes de transportadores ABC, com o objetivo de tornar a anotação de genes mais rápida e identificar qualquer gene alvo em um genoma já montado. Usa Pearl e Bioppearl, porém a comunidade do Python e Biopython é muito maior em relação a resolução de problemas e implementação de novas funcionalidades. O TGFAM-FINDER foi desenvolvido para funcionar no ambiente do sistema operacional Linux. Para um funcionamento adequado requer a instalação dos seguintes programas: Perl 5.6.1 ou versão superior, Bioperl, YAML (módulo perl), Bowtie2-2.3.1, HMMER-3.1b2, BLAST 2.6.0+, InterproScan-5.22-61.0, Exonerate-2.2.0, Blat v35, Tophat-2.1.1 e Cufflinks-2.2.1, Augustus-3.2.3, Scipio-1.4 e ClustalW-2.12.

As tecnologias predominantes identificadas utilizam Python e Biopython, além do Blast para o alinhamento de sequências. É crucial destacar que não foi encontrado nenhum outro *software* com características semelhantes a esta pesquisa, seja em termos de armazenamento de dados ou na geração de novos dados na área de micorremediação de plásticos.

3 METODOLOGIAS

Neste capítulo, serão apresentadas em detalhes as metodologias adotadas para a condução desta pesquisa, englobando tanto a metodologia da pesquisa em si quanto a metodologia do desenvolvimento do método computacional. A descrição detalhada dessas metodologias permitirá uma compreensão clara e transparente dos procedimentos adotados, garantindo a validade e confiabilidade dos resultados obtidos.

3.1 METODOLOGIA DE PESQUISA

A metodologia adotada neste trabalho segue uma abordagem de pesquisa conhecida como *Design Science Research* (DSR) (GOECKS et al., 2021), que se baseia no desenvolvimento e avaliação de artefatos para resolver problemas específicos. Essa abordagem combina aspectos de pesquisa científica e desenvolvimento prático, buscando produzir conhecimento aplicado e soluções tangíveis.

Em relação à natureza, a pesquisa segue uma abordagem aplicada, pois visa criar um método computacional baseado em análise de sequências para a mineração de genomas. O objetivo principal é desenvolver um artefato que possa contribuir para a extração de informações valiosas dos genomas de organismos, proporcionando *insights* e hipóteses para pesquisas futuras na área de bioinformática.

Os objetivos da pesquisa são múltiplos e interligados. Primeiramente, busca-se compreender os desafios e demandas da área de mineração de genomas, identificando lacunas e problemas a serem abordados. Em seguida, o objetivo é projetar e desenvolver um método computacional eficiente, capaz de analisar sequências genômicas e extrair informações relevantes, como a identificação de proteínas com potencial para degradação de plásticos por fungos. Além disso, pretende-se avaliar a eficácia e a aplicabilidade do método desenvolvido testando com pelo menos cinco genomas de fungos diferentes.

Os procedimentos adotados nesta pesquisa seguem a metodologia do DSR, isso envolve a identificação clara do problema a ser abordado, a definição dos objetivos do artefato a ser desenvolvido, o processo de design e implementação do método computacional,

a avaliação da eficácia do artefato e, por fim, a comunicação dos resultados obtidos. Durante o processo de desenvolvimento do método, serão empregados princípios e técnicas de bioinformática, como algoritmos de alinhamento de sequências e análise de dados genômicos.

Em suma, a metodologia da pesquisa combina uma abordagem aplicada, visando a criação de um método computacional, com o *framework* do DSR. Essa abordagem permite uma investigação sistemática e orientada para a solução de problemas, possibilitando o desenvolvimento de um artefato que atenda às necessidades dos pesquisadores na área de bioinformática e contribua para a mineração eficiente de genomas.

No contexto deste projeto o DSR pode ser adaptado da seguinte maneira:

- **Identificação do problema:** A necessidade de um método computacional eficiente para a encontrar potenciais enzimas para a degradação de plástico em genomas fúngicos.
- **Definição dos objetivos do artefato:** Automatizar algumas etapas do processo de mineração e fornecer uma interface para os pesquisadores.
- **Desenvolvimento do artefato:** Projetar e desenvolver o método computacional para a mineração de genomas. Isso envolverá a seleção de algoritmos, a implementação de funcionalidades específicas e a criação de uma arquitetura que permita o processamento eficiente das sequências genômicas.
- **Avaliação do artefato:** Aplicar testes para verificar a funcionalidade do método desenvolvido.
- **Refinamento do artefato:** Refinar o artefato para melhorar sua funcionalidade. Isso pode envolver ajustes nos algoritmos, otimização de recursos computacionais ou melhorias na interface do usuário.
- **Comunicação dos resultados:** Documentar e comunicar sobre os resultados da pesquisa, descrevendo o artefato desenvolvido, as melhorias realizadas e os *insights* obtidos durante o processo.

3.2 METODOLOGIA DE DESENVOLVIMENTO

A metodologia de desenvolvimento adotada neste trabalho é baseada em gestão ágil, utilizando como *framework* os princípios de *OKRs* (*Objectives and Key Results*),

Scrum e *Kanban*. A gestão ágil proporciona um planejamento estratégico orientado por objetivos específicos do projeto e pela medição do seu sucesso. Dessa forma, é possível estabelecer a direção que irá guiar o projeto ao longo do tempo.

O *Scrum* é empregado para realizar o trabalho operacional de forma planejada e eficiente, visando alcançar os objetivos definidos pelos *OKRs*. O trabalho operacional, relacionado à implementação de métodos computacionais e ferramentas de software, é dividido em ciclos chamados de *sprints*. Cada tarefa dentro das *sprints* segue um fluxo de trabalho composto por três etapas: "A fazer", "Em andamento" e "Concluído". Além disso, os papéis no *Scrum* são bem distribuídos e alinhados com a multidisciplinaridade de cada membro da equipe, sendo destacadas as reuniões semanais realizadas para gestão e controle dos projetos.

A metodologia *Kanban*, adaptada ao *Scrum* e baseada em uma abordagem japonesa, é utilizada para gerenciar e controlar as atividades. A ferramenta online *Trello* é empregada para escrever e organizar todas as etapas e tarefas necessárias em *sprints* de uma semana de duração.

Além disso, a gerência de configuração é essencial para controlar e acompanhar as alterações e versões do código fonte do projeto e para fazer isso será usado o Github. O Github é uma plataforma de hospedagem de código-fonte e arquivos com controle de versão usando o Git. Ele permite que programadores, utilitários ou qualquer usuário cadastrado na plataforma contribuam em projetos privados e/ou Open Source de qualquer lugar do mundo.

Em resumo, o Github é um serviço baseado em nuvem que hospeda um sistema de controle de versão (VCS) chamado Git. Ele permite que os desenvolvedores colaborem e façam mudanças em projetos compartilhados enquanto mantêm um registro detalhado do seu progresso. O Github é utilizado como uma plataforma de gerenciamento de repositórios, permitindo que os membros da equipe colaborem no desenvolvimento do software, registrem alterações, resolvam conflitos e tenham um histórico detalhado das modificações realizadas ao longo do tempo.

4 RESULTADOS

Neste capítulo, serão apresentados os resultados alcançados ao longo do desenvolvimento dessa pesquisa, enfocando o Banco de Dados de Enzimas, o Método Computacional e a Aplicação Web.

4.1 BANCO DE DADOS DE ENZIMAS

Foi de importância fundamental estabelecer um banco de dados dedicado ao armazenamento das enzimas com atividade comprovada na degradação de plásticos, juntamente com seus metadados. Nesta seção, iremos explorar em detalhes o processo de modelagem desse banco de dados, bem como apresentar uma visão abrangente das enzimas que foram incorporadas até a data atual desta pesquisa (out/2023).

4.1.1 Modelagem do banco de dados

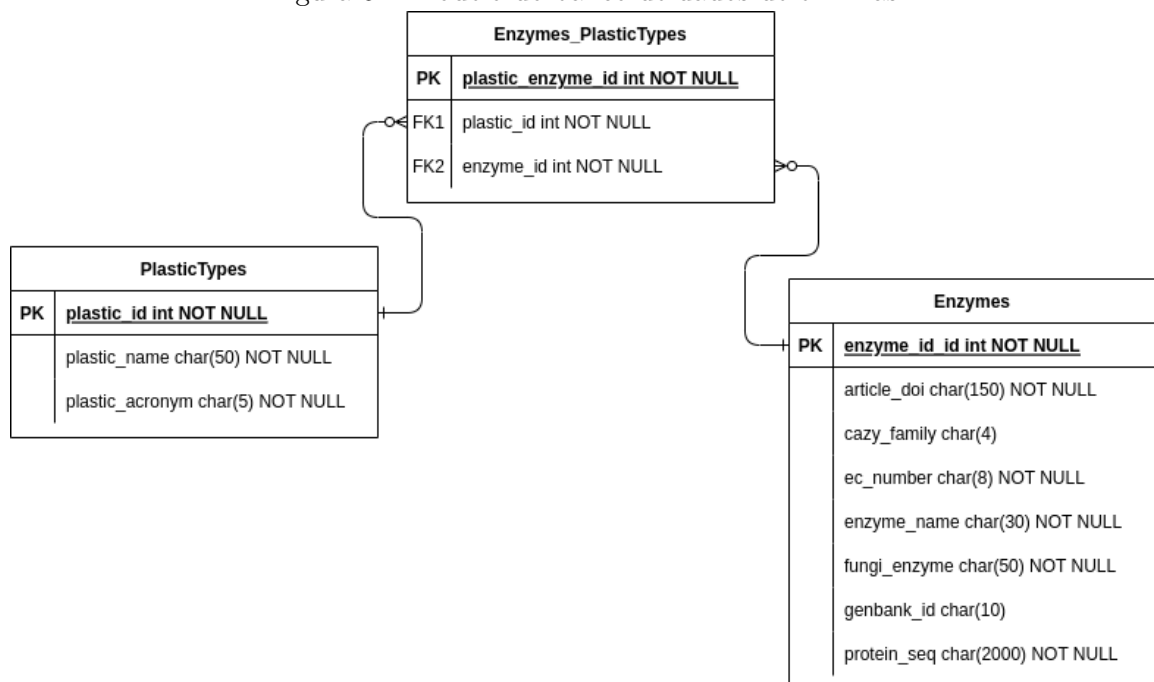
Durante a condução da pesquisa, alguns dados foram considerados mais relevantes e fundamentais para o armazenamento deste conjunto de informações. O banco de dados foi compilado manualmente após uma revisão sistemática da literatura, com foco em identificar enzimas cuja atividade na degradação de plásticos já havia sido comprovada. Informações adicionais foram obtidas nos sites www.cazy.org, que classifica enzimas (CAZymes), <https://www.ncbi.nlm.nih.gov/genbank/> e <https://www.uniprot.org>, onde as sequências de aminoácidos estão disponíveis. Na tabela *Enzymes*, foram incluídas as seguintes informações:

- DOI do artigo: Isso permite a localização de fontes adicionais para obter informações detalhadas sobre a descoberta da enzima.
- CAZY: A classificação na família CAZY à qual a enzima pertence, ajudando na categorização.
- Número EC: O número de classificação da Comissão de Enzimas (EC) que caracteriza a função da enzima.
- Nome da enzima: Para identificação e referência.

- Nome do fungo: O fungo onde a enzima foi originalmente encontrada.
- Número de acesso do GenBank: Quando disponível, esse código facilita o acesso direto à sequência genética da enzima.
- Sequência de aminoácidos: A composição da enzima em termos de sequência de aminoácidos.

Esses dados se tornaram essenciais, não apenas como um meio de comparação e validação com as enzimas a serem testadas na ferramenta, mas também para enriquecer o banco de dados. Além da tabela que abriga metadados sobre as enzimas, o banco de dados inclui outras duas tabelas complementares. Uma delas lista uma variedade de tipos de plástico reconhecidos atualmente, enquanto a terceira tabela estabelece a relação entre as enzimas e o tipo de plástico que são capazes de degradar. Com essas três tabelas interconectadas, o banco de dados se torna um recurso valioso e completo, como mostrado na Figura 6.

Figura 6 – Modelo do banco de dados de enzimas



Fonte: o autor

4.1.2 Enzimas coletadas para o banco de dados

A partir da revisão de artigos científicos foram encontradas uma relação de enzimas que degradam plásticos. Contudo esses artigos não tinham a caracterização

molecular, não identificavam quais vias metabólicas nem quais as sequências de proteínas correspondentes à essas enzimas. Nesse contexto foi fundamental buscar as informações necessárias para tornar o banco de metadados completo, foram realizadas pesquisas adicionais para encontrar as famílias CAZy, os números EC dessas enzimas, obtidos no site (CAZY, 2023) e as sequências de proteínas obtidas nos sites (GENBANK, 2023) e (UNIPROT, 2023). Atualmente, as enzimas da tabela 2 estão armazenadas no banco de dados:

Tabela 2 – Enzimas incluídas no banco de dados.

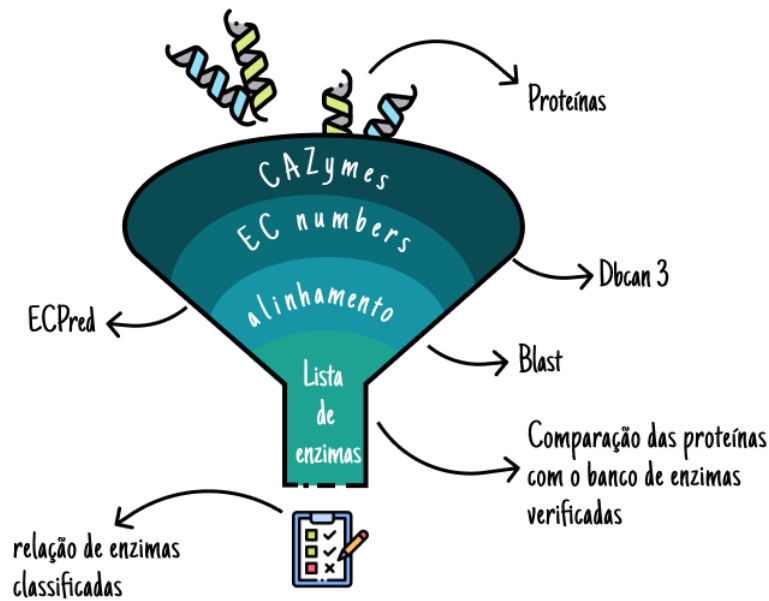
-	Nome do fungo	Enzima	Código EC	Família Cazy
1	<i>Pleurotus sp. 'Florida'</i>	<i>manganese peroxidase</i>	1.11.1.13	AA2
2	<i>Trametes versicolor</i>	<i>lignin peroxidase</i>	1.11.1.14	AA2
3	<i>Pleurotus eryngii</i>	<i>versatile peroxidase</i>	1.11.1.16	AA2
4	<i>Aspergillus oryzae</i>	<i>triacylglycerol lipase</i>	3.1.1.3	CE5
5	<i>Fusarium solani</i>	<i>cutinase</i>	3.1.1.74	CE5
6	<i>Bjerkandera adusta</i>	<i>versatile peroxidase</i>	1.11.1.16	AA2
7	<i>Aspergillus terreus</i>	<i>esterase</i>	3.1.1.-	CE3
8	<i>Aspergillus flavus</i>	<i>laccase</i>	1.10.3.2	AA1
9	<i>Phanerodontia chrysosporium</i>	<i>manganese peroxidase I precursor</i>	1.11.1.13	AA2
10	<i>Pleurotus ostreatus</i>	<i>manganese peroxidase</i>	1.11.1.13	AA2
11	<i>Trametes versicolor</i>	<i>laccase3</i>	1.10.3.2	AA1
12	<i>Humicola insolens</i>	<i>cutinase</i>	3.1.1.74	CE5
13	<i>Fusarium oxysporum f. sp. cepae</i>	<i>Cutinase 3</i>	3.1.1.74	CE5
14	<i>Pleurotus ostreatus</i>	<i>laccase</i>	1.10.3.2	AA1
15	<i>Phanerodontia chrysosporium</i>	<i>lignin peroxidase</i>	1.11.1.14	AA2

Fonte: Tabela: O autor, Dados: (TEMPORITI et al., 2022), (CAZY, 2023), (GENBANK, 2023), (UNIPROT, 2023)

4.2 PLASTICOME: MÉTODO COMPUTACIONAL

O Plasticome é uma ferramenta que implementa um método computacional que minera um genoma de um fungo buscando enzimas que degradam plástico, como o processo consta com várias fases foi desenvolvido na arquitetura de pipeline (vide seção 2.2.1.1). A pipeline escolhida para compor este projeto é uma sequência de etapas que visa realizar a análise computacional das enzimas envolvidas no processo de degradação de plásticos. Essa pipeline, representada graficamente pela figura 7 é composta pelos seguintes passos:

Figura 7 – Representação gráfica para os passos da pipeline definida



Fonte: O autor

1. Um arquivo em formato fasta com todas as proteínas anotadas do organismo escolhido pelo usuário é baixado do GenBank, a figura 8 mostra um exemplo de um arquivo fasta de proteínas.

Figura 8 – Representação de três proteínas em um arquivo fasta

```

1 >KAF7722121.1 Uncharacterized protein PECH_005776 [Penicillium ucsense]
2 MEIQQRLLDQKELCEGLVQSKPAVEEQSLNSMEIMKTLVSHQNVDSNMRRAIVDGIRGYKSLLRSDGRSPISATTPELL
3 TRGPDPAAGSLHEDFTTNGPQGS LHCPFSKPTTPRPGSDMASGRRDDGPKILIDDT CGHDDLDP IKADQEERRSSTAPSVG
4 RSSQGHCPVSRCPTRYLDKHSPKEIAEYVERHKHEIPRSHAICVKRYQRPQNMRLDAKYGGLINMIRGLSAKHQAFPL
5 EREANEDGEVDEQSGHSSSPERVEKWAEEDVPVTPMPPTQPADERESRFRPLREVVRVGESPSRPWGPVPIAEQPPAPS
6 SYFPSSNAPTPEPTTAAPSISSPHQPPAKPAGRCFPGHDAPKQKPVASPPAPAPAPASAPPAPAAHLESPWSNW
7 GNVDLTNTGQATVGQGGTSRIENGPRIDGQSPVSHVTFNGPVFFGSAEQTASLLQQLAQNKDPSQG
8 >KAF7722122.1 L-PSP endoribonuclease family protein, partial [Penicillium ucsense]
9 MSAHMSAVTAKDACPPAGPYSQATRANGQIFVSGQIPADSTGALVQGTIGELTQACNNIDAILKAAGSEVSKIVKVNWF
10 LTDMANFAEMNATYEKFFVHKPARSCVAVHQLPKGVLVEIEICIAL
11 >KAF7722162.1 Uncharacterized protein PECH_005542 [Penicillium ucsense]
12 MLTHFQEEAEQIYKNGQKAVVPAGNAGRTGAGIYALRGFRQWNEVSTMWDCAMTIDSDVWNGWNKIWLPRFFEPEDK
13 EKDPQTCKPRDLCGMTPIQKDRARFLQYIDPSYTVDNVIFSNVSHKDKVQCIYIPNGLVKTNIWLTPCTERLAKDSE
14 RLGVFEFGTAEWDLTSLPGWGLGVLPQ
  
```

Fonte: O Autor

Cada ‘>’ representa o início de uma anotação de proteína, e na linha logo abaixo são descritos todos os aminoácidos que formam essa proteína.

2. Nesta etapa, a ferramenta DBCAN é utilizada para identificar quais enzimas do arquivo de proteínas, extraído na etapa anterior, são ativas em carboidratos, assim como determinar a qual família CAZy essas enzimas pertencem, a figura 9 mostra o

formato do resultado da identificação de cada proteína.

Figura 9 – Exemplo de saída do dbcan, em um arquivo separado por tabulação

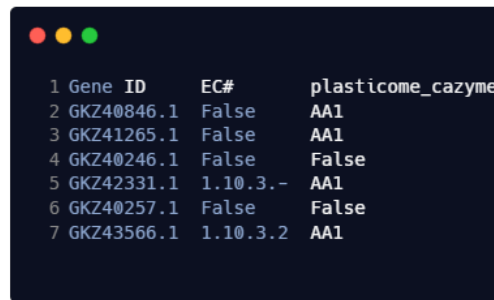
1	Gene ID	EC#	HMMER	eCAMI	DIAMOND	#ofTools	
2	KAF7722350.1	-	GH71(22-406)+CBM24(448-522)+CBM24(541-617)	CBM24+GH71	GH71	3	
3	TESTE350.1	-	GH1(22-406)+CE5(448-522)+CBM24(541-617)	CBM24+GH71	GH71	3	
4	22222350.1	-	GH1(22-406)+CE2(448-522)+CBM24(541-617)	CBM24+GH71	AA2	3	
5	PEUE350.1	-	GH1(22-406)+CE1(448-522)+CBM24(541-617)	CBM24+GH71	AA3	3	

Fonte: O Autor

O cabeçalho dessa saída representa os seguintes dados:

- Gene ID*: O id da proteína analisada;
 - EC*: Um possível número EC para a proteína;
 - HMMER*: Resultado da consulta contra o banco de dados dbCAN HMM (modelo oculto de Markov), nesse caso interpretando a primeira linha a proteína foi classificada pelo HMMER como GH71 da sua posição 22 à 406, como CBM24 da sua posição 448 à 522 e também da posição 541 à 617;
 - eCAMI*: resultado da análise de acordo com a ferramenta eCAMI do dbCAN, que classificou a primeira proteína como CBM24+GH71;
 - DIAMOND*: Outra ferramenta usada pelo dbCAN que classificou toda a proteína da primeira linha como sendo da família GH71;
 - ofTools*: Representa a quantidade de ferramentas que retornam um resultado, o valor 3 significa que todas as 3 ferramentas retornaram um resultado.
3. Após a anotação das CAZymes, o arquivo de proteínas passa por um processo de filtragem, restando apenas aquelas que foram identificadas com alguma família CAZy presente no banco de metadados do plasticome, na figura 10, é possível ver quais dados restam após o filtro, também é possível identificar na coluna 'plasticome_cazyme' quais enzimas não possuem família cazy definida no banco de metadados, marcadas com o valor booleano "False" e quais possuem, com a respectiva família CAZy.

Figura 10 – Exemplo de saída após filtro de CAZyS

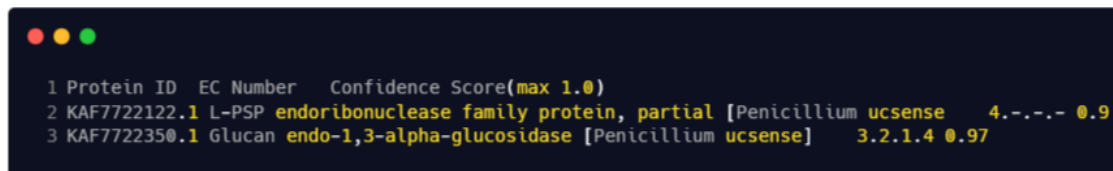


1	Gene ID	EC#	plasticome_cazyme
2	GKZ40846.1	False	AA1
3	GKZ41265.1	False	AA1
4	GKZ40246.1	False	False
5	GKZ42331.1	1.10.3.-	AA1
6	GKZ40257.1	False	False
7	GKZ43566.1	1.10.3.2	AA1

Fonte: O Autor

4. Após a identificação das famílias CAZy, a ferramenta ECPred é empregada para determinar os números EC correspondentes a cada enzima restante no arquivo de proteínas, Na figura 11 é possível ver como o ECPred apresenta os resultados.

Figura 11 – Exemplo de saída da fase do ECPred, em um arquivo separado por tabulação



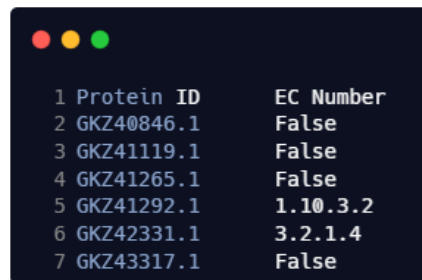
1	Protein ID	EC Number	Confidence Score(max 1.0)
2	KAF7722122.1	L-PSP endoribonuclease family protein, partial [Penicillium ucsense]	4.-.-.- 0.9
3	KAF7722350.1	Glucan endo-1,3-alpha-glucosidase [Penicillium ucsense]	3.2.1.4 0.97

Fonte: O Autor

Para interpretar corretamente os dados da saída do ECPred, leve os seguintes fatos em consideração:

- Protein ID*: Todo o cabeçalho da proteína;
 - EC Number*: O número EC previsto para essa proteína, no contexto da primeira linha onde encontra-se o valor "4.-.-"significa que o ECPred conseguiu encontrar apenas a classe principal e não obteve correspondências para as subclasses;
 - Confidence Score*: O percentual da confiança com a predição do EC, o valor pode variar de 0 à 1.0.
5. Assim como as enzimas que possuíam famílias CAZy divergentes do banco de metadados foram descartadas, acontece com enzimas que possuem números EC divergentes, mantendo assim um arquivo com apenas enzimas que possuem ambos (família CAZy e número EC) presentes no banco, o resultado fica como na figura 12.

Figura 12 – Exemplo de saída do filtro de números EC, em um arquivo separado por tabulação



1 Protein ID	EC Number
2 GKZ40846.1	False
3 GKZ41119.1	False
4 GKZ41265.1	False
5 GKZ41292.1	1.10.3.2
6 GKZ42331.1	3.2.1.4
7 GKZ43317.1	False

Fonte: O Autor

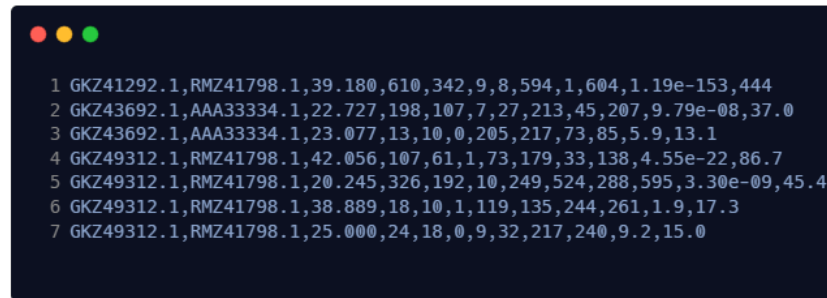
6. Após essa segunda filtragem, o arquivo contendo as enzimas restantes é fragmentado para realizar um alinhamento de um para um com a enzima do banco de dados que compartilha a mesma família CAZy e o mesmo número EC, utilizando a ferramenta BLAST, o BLAST não retorna um cabeçalho para seus resultados (figura 13), porém as colunas devem ser interpretadas da seguinte forma:

- QUERY ID*: O id da proteína que foi consultada,
- REF ID*: O id da proteína de referência,
- IDENTITY (%)*: A porcentagem de identidade entre as duas proteínas,
- LENGTH*: Indica tamanho do alinhamento,
- MISMATCHES*: Indica o número de nucleotídeos/aminoácidos diferentes em um alinhamento,
- GAP OPENS*: Representa a quantidade de lacunas,
- Q. START*: Posição de início do alinhamento na sequência de consulta,
- Q. END*: Posição do término do alinhamento na sequência de consulta,
- S. START*: Posição de início do alinhamento na sequência de referência,
- S. END*: Posição de término do alinhamento na sequência de referência,
- E-VALUE*: Medida da probabilidade de que um alinhamento observado seja devido ao acaso,
- BIT SCORE*: Valor de pontuação, que representa a qualidade do alinhamento.

O BLAST em alguns alinhamentos de pares calcula diferentes alinhamentos que possuem diferentes similaridades (c) e diferentes bit-score (l), por representar a qualidade do alinhamento nesses casos foram escolhidos os alinhamentos com maior

bit-score (NCBI. . . , 2023).

Figura 13 – Exemplo de saída do alinhamento com Blast



Fonte: O Autor

7. Por fim, depois das análises, filtrações e alinhamento das enzimas, realiza-se uma comparação de classificação com o banco de metadados para determinar os tipos de plásticos para os quais cada enzima possui potencial de degradação. Os resultados são apresentados em dois formatos (figura 14 e tabela 3): primeiro, um gráfico que relaciona cada enzima potencial identificada no genoma com os tipos de plásticos que pode degradar; e segundo, uma tabela que classifica o percentual de similaridade entre a enzima consultada e a enzima similar encontrada no banco de metadados.

Figura 14 – Gráfico enviado ao usuário relacionando enzimas e tipos de plástico



Fonte: O Autor

Tabela 3 – Resultado 2: relação de similaridade entre enzimas encontradas no genoma e enzimas na base do Plasticome.

-	Enzima consultada	Enzima com atividade comprovada	Similaridade (%)
1	PIL36360.1 transporter [<i>Ganoderma sinense</i> ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [<i>Aspergillus flavus</i>]	22.951
2	PIL23007.1 hypothetical protein [<i>Ganoderma sinense</i> ZZ0214-1]	GFF16684.1 carbohydrate esterase family 3 protein [<i>Aspergillus terreus</i>]	100
3	PIL23727.1 transporter [<i>Ganoderma sinense</i> ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [<i>Aspergillus flavus</i>]	34.444
4	PIL29303.1 transporter [<i>Ganoderma sinense</i> ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [<i>Aspergillus flavus</i>]	31.111
5	PIL30321.1 hypothetical protein [<i>Ganoderma sinense</i> ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [<i>Aspergillus flavus</i>]	32.967
6	PIL30422.1 transporter [<i>Ganoderma sinense</i> ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [<i>Aspergillus flavus</i>]	24.04
7	PIL30424.1 transporter [<i>Ganoderma sinense</i> ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [<i>Aspergillus flavus</i>]	31.098
8	PIL31400.1 transporter [<i>Ganoderma sinense</i> ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [<i>Aspergillus flavus</i>]	32.778
9	PIL33565.1 transporter [<i>Ganoderma sinense</i> ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [<i>Aspergillus flavus</i>]	32.804
10	PIL36359.1 transporter [<i>Ganoderma sinense</i> ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [<i>Aspergillus flavus</i>]	25.043

Fonte: O autor

A escolha dessa pipeline visa integrar diferentes ferramentas e técnicas computacionais para realizar uma análise abrangente e detalhada das enzimas envolvidas na degradação de plásticos, do fungo alvo. Essa abordagem permite uma melhor compreensão dos processos de degradação e contribui para o desenvolvimento de estratégias mais eficientes para a biodegradação dos plásticos.

4.3 APLICAÇÃO WEB

Neste capítulo, exploraremos em detalhes a aplicação web "Plasticome", incluindo sua arquitetura, funcionalidades-chave e como ela pode ser uma ferramenta valiosa para a pesquisa em bioinformática. Além disso, será evidenciada a verificação de funcionalidade e caracterização da ferramenta.

4.3.1 Arquitetura da Aplicação

A aplicação possui uma interface que pode ser visualizada na figura 15 com três campos. O usuário deve fornecer seu nome, e-mail para o envio dos resultados e o número de acesso do GenBank correspondente ao genoma escolhido para análise. É fundamental observar que, neste momento, a aplicação é funcional apenas para genomas registrados no GenBank que tenham suas sequências de proteínas devidamente anotadas, caso seja enviado um número de acesso de um genoma que não possui anotação de proteínas, a análise não será executada.

Além disso, o Plasticome está comprometido com a privacidade dos usuários. As informações pessoais fornecidas, como nome e e-mail, são usadas apenas para o propósito de enviar os resultados da análise e não são armazenadas ou compartilhadas com terceiros. Essa plataforma fornece acesso direto por sessão sem precisar cadastrar uma conta com login e senha.

Figura 15 – Interface da ferramenta

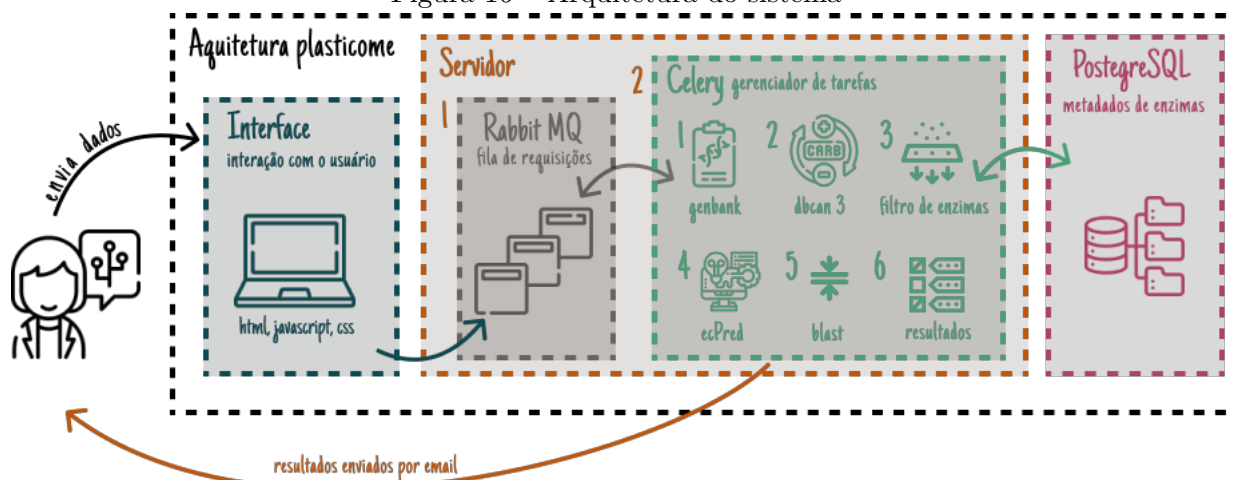


Fonte: o autor

Após o envio dos dados, o usuário recebe uma mensagem de alerta na tela, informando que sua análise está em andamento e que o resultado será enviado por e-mail. A requisição é recebida pelo servidor e entra em uma fila orquestrada pelo RabbitMQ, dando início à pipeline. Esta é mediada pelo Celery, que cria uma cadeia de tarefas subsequentes.

A arquitetura completa do ferramenta está representada na figura 16.

Figura 16 – Arquitetura do sistema



Fonte: O autor

A primeira tarefa é o download do arquivo fasta de proteínas do genoma escolhido do genbank (figura 16, setor servidor, bloco 2, passo 1). Em seguida, a análise é realizada pelo dbCAN, através de um container Docker que utiliza a imagem do projeto autônomo do servidor web dbCAN3, *run_dbcan* (acessível em: https://github.com/linnabrown/run_dbcan).

O passo seguinte é a filtragem do arquivo fasta com base no resultado do dbCAN, mantendo apenas as proteínas que foram classificadas pelo dbCAN como pertencentes às famílias CAZy cadastradas no banco de dados. Este arquivo fasta filtrado é enviado ao container do ECPred (acessível em: <https://github.com/cansyl/ECPred>), que teve sua imagem Docker criada e publicada durante a execução deste projeto (acessível em: <https://hub.docker.com/u/blueevee>).

Após o recebimento do resultado do ECPred, é realizado outro filtro no arquivo, desta vez para manter apenas as enzimas com números EC que existem no banco de dados. Em seguida, é realizado um alinhamento de sequência usando o BLAST para verificar a similaridade entre a enzima do fungo consultado e uma enzima com atividade comprovada na degradação de plásticos que possua o mesmo número EC e família CAZy.

Por fim, é realizada uma comparação das enzimas do banco de dados e seus respectivos números EC e famílias CAZy com as classificações disponibilizadas pelo dbCAN e ECPred para identificar quais tipos de plásticos essas enzimas têm a possibilidade de degradar.

Caso não exista nenhuma correspondência, é enviado um e-mail para o usuário informando que, no momento, não foram identificadas enzimas com possibilidade para degradação de plásticos. No entanto, isso pode mudar no futuro, tendo em vista que a base do Plasticome deverá aumentar e, com isso, ampliar a relação com fungos diferentes.

Caso existam enzimas identificadas, são gerados dois resultados: uma imagem que possui um gráfico relacionando cada enzima com os tipos de plástico que pode degradar e uma tabela com a relação da similaridade encontrada no alinhamento feito pelo BLAST. Esses resultados podem ser usados para certificar e aumentar a confiabilidade de que aquela enzima realmente consiga degradar aquele tipo de plástico. Contudo podem existir enzimas com alta capacidade de degradação que ainda não tenham sido identificadas e armazenadas na nossa base.

4.3.2 Verificação de funcionalidade

A verificação da aplicação web envolveu a definição de um protocolo de testes para verificar o desempenho e a funcionalidade da aplicação. Esse protocolo foi dividido em três tarefas principais:

1. Teste de funcionalidade do banco de dados de genes com atividade comprovada na degradação de plásticos.
2. Definição de testes utilizando pelo menos 5 genomas de fungos: Nessa etapa, foram definidos cinco genomas de diferentes ordens de fungos que possuem anotação de proteínas no Genbank. As ordens dos fungos são categorias taxonômicas que agrupam diferentes famílias de fungos com características semelhantes, existem diversas delas, porém cinco ordens distintas foram escolhidas arbitrariamente com uma espécie de fungo por ordem, os fungos escolhidos podem ser visualizados na tabela 4.

Tabela 4 – Espécies de fungos escolhidos para fazer os testes no Plasticome

Ordem do fungo	Espécie escolhida	Número de acesso do Genbank
<i>Sordariales</i>	<i>Neurospora tetrasperma</i> FGSC 2508	GCA_000213175.1
<i>Eurotiales</i>	<i>Aspergillus brasiliensis</i> IFM 66951	GCA_027924065.1
<i>Hipocreales</i>	<i>Fusarium oxysporum</i> Fo47	GCA_013085055.1
<i>Polyporales</i>	<i>Ganoderma sinense</i> ZZ0214-1	GCA_002760635.1
<i>Agaricales</i>	<i>Pleurotus pulmonarius</i> PM_ss5	GCA_012979565.1

Fonte: O autor

3. Execução e Organização dos Testes: Uma vez estabelecido o protocolo de testes, procedeu-se à sua execução conforme planejado. Cada análise resultou em dois arquivos para os fungos que apresentaram enzimas com potencial de degradação de plásticos. Adicionalmente, registrou-se o tempo decorrido em cada fase para todas as requisições. Os resultados detalhados dos testes estão disponíveis abaixo.

4.3.2.1 Resultados da verificação

Nesta subseção, estão disponíveis tabelas e figuras que documentam os resultados dos testes executados na ferramenta web Plasticome. Na Tabela 5, é possível verificar a quantidade de proteínas anotadas para o organismo usado em cada requisição, o tempo (em segundos) que cada requisição levou em cada etapa do pipeline, o tempo total em horas, além de identificar quantas enzimas o fungo consultado possui com potencial para a degradação de plásticos. Os experimentos foram realizados em um notebook com sistema operacional windows versão 10, 16GB de RAM, Processador Intel i7-7500U 2.70GHz, portanto o tempo pode variar de acordo com o ambiente utilizado para execução.

Nas Figuras 17, 18, 19, 20 e 21, apresentam-se os gráficos correspondentes a cada consulta. Cada gráfico exibe uma marcação em forma de círculo colorido, indicando a relação entre a enzima e os tipos de plásticos que ela potencialmente pode degradar. Adicionalmente, um quadro contendo a nomenclatura de todos os tipos de plástico testados durante a consulta está disponível.

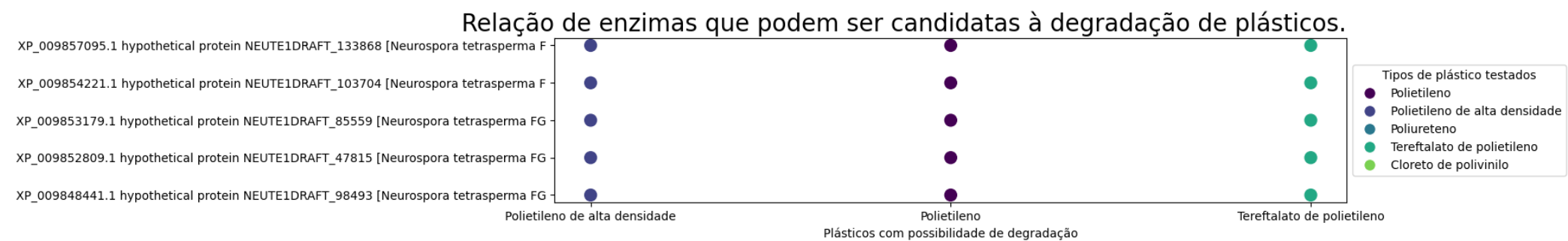
Nas Tabelas 6, 7, 8, 9 e 10, encontram-se os resultados dos alinhamentos realizados pelo BLAST. Algumas enzimas podem possuir mais de um alinhamento quando mais de uma enzima semelhante foi encontrada no banco de dados.

Tabela 5 – Resultado dos testes executados, com tempo de execução em segundos em cada fase.

-	Espécie escolhida	Proteínas	DbCan (s)	Filtro de CAZy (s)	EcPred (s)	Filtro de EC (s)	BLAST (s)	Total (H:m)	Enzim
1	GCA_000213175.1	10.380	4837.109	5.875	1343.766	0.160	14.937	1:43	
2	GCA_027924065.1	11.888	7627.530	6.406	1955.483	0.112	14.590	2:40	
3	GCA_013085055.1	16.202	7848.0626	5.125	2752.1098	0.309	15.046	2:56	
4	GCA_002760635.1	15.478	6425.047	7.172	1992.687	0.977	25.797	2:21	
5	GCA_012979565.1	11.873	5944.094	6.530	1508.531	0.030	16.672	2:04	

Fonte: O autor

Figura 17 – **Consulta 1:** Resultado gráfico com potenciais enzimas encontradas no *Neurospora tetrasperma* FGSC 2508.



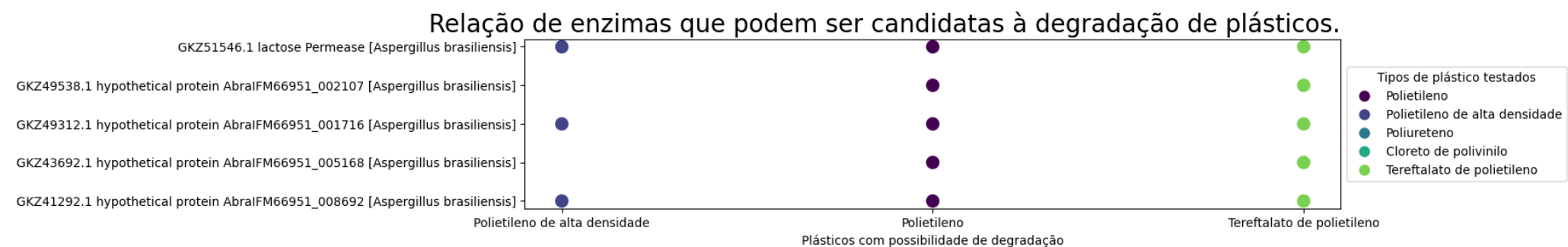
Fonte: O autor

Tabela 6 – **Consulta 1:** Similaridade das enzimas encontradas no *Neurospora tetrasperma* FGSC 2508 com o banco de dados Plasticome.

Enzima consultada	Enzima com atividade comprovada	Similaridade (%)
XP_009848441.1 hypothetical protein NEUTE1DRAFT_98493 [Neurospora tetrasperma FGSC 2508]	AAR21094.1 laccase [Pleurotus ostreatus]	31,515
XP_009848441.1 hypothetical protein NEUTE1DRAFT_98493 [Neurospora tetrasperma FGSC 2508]	BAD98307.1 laccase3 [Trametes versicolor]	32,319
XP_009848441.1 hypothetical protein NEUTE1DRAFT_98493 [Neurospora tetrasperma FGSC 2508]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	30,263
XP_009852809.1 hypothetical protein NEUTE1DRAFT_47815 [Neurospora tetrasperma FGSC 2508]	AAR21094.1 laccase [Pleurotus ostreatus]	33,013
XP_009852809.1 hypothetical protein NEUTE1DRAFT_47815 [Neurospora tetrasperma FGSC 2508]	BAD98307.1 laccase3 [Trametes versicolor]	31,927
XP_009852809.1 hypothetical protein NEUTE1DRAFT_47815 [Neurospora tetrasperma FGSC 2508]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	22,648
XP_009853179.1 hypothetical protein NEUTE1DRAFT_85559 [Neurospora tetrasperma FGSC 2508]	AAR21094.1 laccase [Pleurotus ostreatus]	29,981
XP_009853179.1 hypothetical protein NEUTE1DRAFT_85559 [Neurospora tetrasperma FGSC 2508]	BAD98307.1 laccase3 [Trametes versicolor]	31,619
XP_009853179.1 hypothetical protein NEUTE1DRAFT_85559 [Neurospora tetrasperma FGSC 2508]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	40,323
XP_009854221.1 hypothetical protein NEUTE1DRAFT_103704 [Neurospora tetrasperma FGSC 2508]	AAR21094.1 laccase [Pleurotus ostreatus]	28,624
XP_009854221.1 hypothetical protein NEUTE1DRAFT_103704 [Neurospora tetrasperma FGSC 2508]	BAD98307.1 laccase3 [Trametes versicolor]	29,026
XP_009854221.1 hypothetical protein NEUTE1DRAFT_103704 [Neurospora tetrasperma FGSC 2508]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	38,835
XP_009857095.1 hypothetical protein NEUTE1DRAFT_133868 [Neurospora tetrasperma FGSC 2508]	AAR21094.1 laccase [Pleurotus ostreatus]	28,731
XP_009857095.1 hypothetical protein NEUTE1DRAFT_133868 [Neurospora tetrasperma FGSC 2508]	BAD98307.1 laccase3 [Trametes versicolor]	31,936
XP_009857095.1 hypothetical protein NEUTE1DRAFT_133868 [Neurospora tetrasperma FGSC 2508]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	23,702

Fonte: O autor

Figura 18 – **Consulta 2:** Resultado gráfico com potenciais enzimas encontradas no *Aspergillus brasiliensis* IFM 66951.



Fonte: O autor

Tabela 7 – **Consulta 2:** Similaridade das enzimas encontradas no *Aspergillus brasiliensis* IFM 66951 com o banco de dados Plasticome.

Enzima consultada	Enzima com atividade comprovada	Similaridade (%)
GKZ41292.1 hypothetical protein AbraIFM66951_008692 [Aspergillus brasiliensis]	AAR21094.1 laccase [Pleurotus ostreatus]	24,621
GKZ41292.1 hypothetical protein AbraIFM66951_008692 [Aspergillus brasiliensis]	BAD98307.1 laccase3 [Trametes versicolor]	23,438
GKZ41292.1 hypothetical protein AbraIFM66951_008692 [Aspergillus brasiliensis]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	39,18
GKZ43692.1 hypothetical protein AbraIFM66951_005168 [Aspergillus brasiliensis]	AAA33334.1 cutinase [Fusarium solani]	22,727
GKZ43692.1 hypothetical protein AbraIFM66951_005168 [Aspergillus brasiliensis]	AAE13316.1 Sequence 2 from patent US 5827719	23,256
GKZ43692.1 hypothetical protein AbraIFM66951_005168 [Aspergillus brasiliensis]	RKK11869.1 Cutinase 3 [Fusarium oxysporum f. sp. cepae]	25,234
GKZ49312.1 hypothetical protein AbraIFM66951_001716 [Aspergillus brasiliensis]	AAR21094.1 laccase [Pleurotus ostreatus]	30,181
GKZ49312.1 hypothetical protein AbraIFM66951_001716 [Aspergillus brasiliensis]	BAD98307.1 laccase3 [Trametes versicolor]	31,546
GKZ49312.1 hypothetical protein AbraIFM66951_001716 [Aspergillus brasiliensis]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	42,056
GKZ49538.1 hypothetical protein AbraIFM66951_002107 [Aspergillus brasiliensis]	AAA33334.1 cutinase [Fusarium solani]	48,571
GKZ49538.1 hypothetical protein AbraIFM66951_002107 [Aspergillus brasiliensis]	AAE13316.1 Sequence 2 from patent US 5827719	50,644
GKZ49538.1 hypothetical protein AbraIFM66951_002107 [Aspergillus brasiliensis]	RKK11869.1 Cutinase 3 [Fusarium oxysporum f. sp. cepae]	44,978
GKZ51546.1 lactose Permease [Aspergillus brasiliensis]	AAR21094.1 laccase [Pleurotus ostreatus]	29,135
GKZ51546.1 lactose Permease [Aspergillus brasiliensis]	BAD98307.1 laccase3 [Trametes versicolor]	30,862
GKZ51546.1 lactose Permease [Aspergillus brasiliensis]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	35,795

Fonte: O autor

Figura 19 – **Consulta 3:** Resultado gráfico com potenciais enzimas encontradas no *Fusarium oxysporum* Fo47.



Fonte: O autor

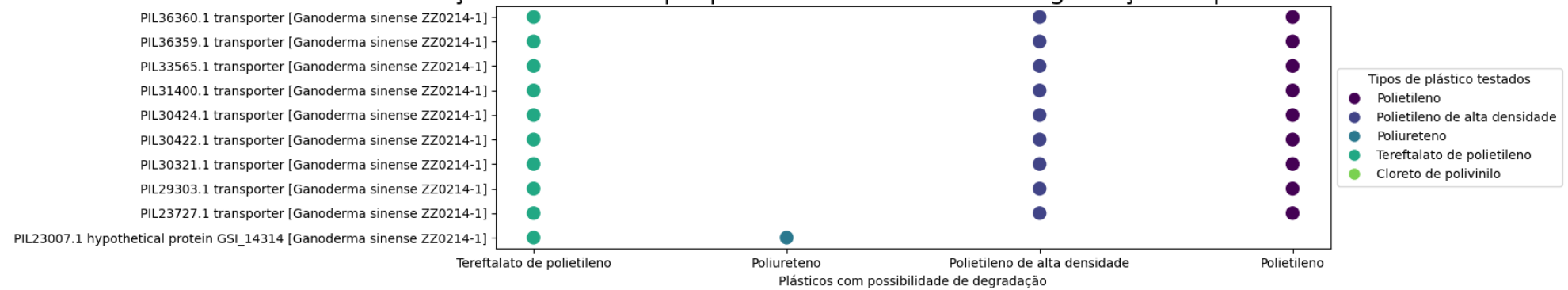
Tabela 8 – **Consulta 3:** Similaridade das enzimas encontradas no *Fusarium oxysporum* Fo47 com o banco de dados Plasticome.

Enzima consultada	Enzima com atividade comprovada	Similaridade (%)
XP_031036799.2 Cupredoxin [Fusarium oxysporum Fo47]	AAR21094.1 laccase [Pleurotus ostreatus]	32,128
XP_031036799.2 Cupredoxin [Fusarium oxysporum Fo47]	BAD98307.1 laccase3 [Trametes versicolor]	36,439
XP_031036799.2 Cupredoxin [Fusarium oxysporum Fo47]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	31,126
XP_031038750.2 cutinase-domain-containing protein [Fusarium oxysporum Fo47]	AAA33334.1 cutinase [Fusarium solani]	73,593
XP_031038750.2 cutinase-domain-containing protein [Fusarium oxysporum Fo47]	AAE13316.1 Sequence 2 from patent US 5827719	53,879
XP_031038750.2 cutinase-domain-containing protein [Fusarium oxysporum Fo47]	RKK11869.1 Cutinase 3 [Fusarium oxysporum f. sp. cepae]	69,697
XP_031041063.2 multicopper oxidase-domain-containing protein [Fusarium oxysporum Fo47]	AAR21094.1 laccase [Pleurotus ostreatus]	28,458
XP_031041063.2 multicopper oxidase-domain-containing protein [Fusarium oxysporum Fo47]	BAD98307.1 laccase3 [Trametes versicolor]	31,238
XP_031041063.2 multicopper oxidase-domain-containing protein [Fusarium oxysporum Fo47]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	24,796
XP_031044377.2 Cupredoxin [Fusarium oxysporum Fo47]	AAR21094.1 laccase [Pleurotus ostreatus]	29,15
XP_031044377.2 Cupredoxin [Fusarium oxysporum Fo47]	BAD98307.1 laccase3 [Trametes versicolor]	32,296
XP_031044377.2 Cupredoxin [Fusarium oxysporum Fo47]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	23,986
XP_054558657.2 cutinase [Fusarium oxysporum Fo47]	GFF16684.1 carbohydrate esterase family 3 protein [Aspergillus terreus]	22,727

Fonte: O autor

Figura 20 – **Consulta 4:** Resultado gráfico com potenciais enzimas encontradas no *Ganoderma sinense* ZZ0214-1.

Relação de enzimas que podem ser candidatas à degradação de plásticos.



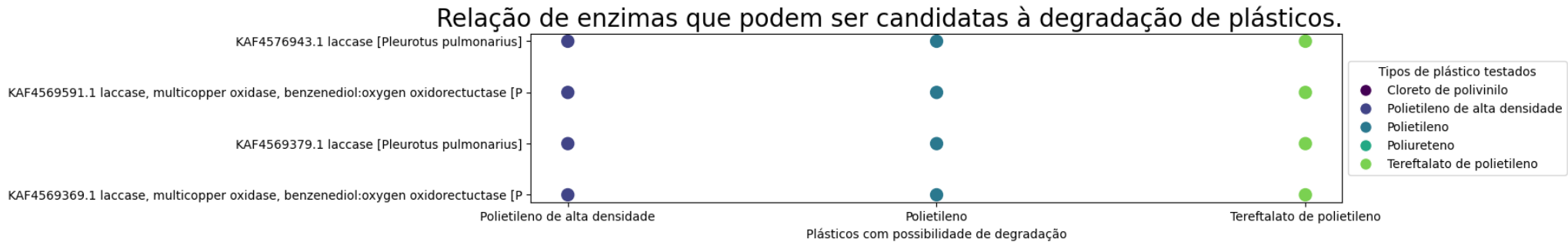
Fonte: O autor

Tabela 9 – **Consulta 4:** Similaridade das enzimas encontradas no *Ganoderma sinense* ZZ0214-1 com o banco de dados Plasticome.

Enzima consultada	Enzima com atividade comprovada	Similaridade (%)
PIL36360.1 transporter [Ganoderma sinense ZZ0214-1]	AAR21094.1 laccase [Pleurotus ostreatus]	59,687
PIL36360.1 transporter [Ganoderma sinense ZZ0214-1]	BAD98307.1 laccase3 [Trametes versicolor]	66.8
PIL36360.1 transporter [Ganoderma sinense ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	22,951
PIL23007.1 hypothetical protein GSI_14314 [Ganoderma sinense ZZ0214-1]	GFF16684.1 carbohydrate esterase family 3 protein [Aspergillus terreus]	100
PIL23727.1 transporter [Ganoderma sinense ZZ0214-1]	AAR21094.1 laccase [Pleurotus ostreatus]	54,739
PIL23727.1 transporter [Ganoderma sinense ZZ0214-1]	BAD98307.1 laccase3 [Trametes versicolor]	62,725
PIL23727.1 transporter [Ganoderma sinense ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	34,444
PIL29303.1 transporter [Ganoderma sinense ZZ0214-1]	AAR21094.1 laccase [Pleurotus ostreatus]	61,961
PIL29303.1 transporter [Ganoderma sinense ZZ0214-1]	BAD98307.1 laccase3 [Trametes versicolor]	73,653
PIL29303.1 transporter [Ganoderma sinense ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	31,111
PIL30321.1 transporter [Ganoderma sinense ZZ0214-1]	AAR21094.1 laccase [Pleurotus ostreatus]	60,314
PIL30321.1 transporter [Ganoderma sinense ZZ0214-1]	BAD98307.1 laccase3 [Trametes versicolor]	62.5
PIL30321.1 transporter [Ganoderma sinense ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	32,967
PIL30422.1 transporter [Ganoderma sinense ZZ0214-1]	AAR21094.1 laccase [Pleurotus ostreatus]	55,166
PIL30422.1 transporter [Ganoderma sinense ZZ0214-1]	BAD98307.1 laccase3 [Trametes versicolor]	59,375
PIL30422.1 transporter [Ganoderma sinense ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	24,04
PIL30424.1 transporter [Ganoderma sinense ZZ0214-1]	AAR21094.1 laccase [Pleurotus ostreatus]	54,403
PIL30424.1 transporter [Ganoderma sinense ZZ0214-1]	BAD98307.1 laccase3 [Trametes versicolor]	59,761
PIL30424.1 transporter [Ganoderma sinense ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	31,098
PIL31400.1 transporter [Ganoderma sinense ZZ0214-1]	AAR21094.1 laccase [Pleurotus ostreatus]	61,942
PIL31400.1 transporter [Ganoderma sinense ZZ0214-1]	BAD98307.1 laccase3 [Trametes versicolor]	67,878
PIL31400.1 transporter [Ganoderma sinense ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	32,778
PIL33565.1 transporter [Ganoderma sinense ZZ0214-1]	AAR21094.1 laccase [Pleurotus ostreatus]	58,414
PIL33565.1 transporter [Ganoderma sinense ZZ0214-1]	BAD98307.1 laccase3 [Trametes versicolor]	59,501
PIL33565.1 transporter [Ganoderma sinense ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	32,804
PIL36359.1 transporter [Ganoderma sinense ZZ0214-1]	AAR21094.1 laccase [Pleurotus ostreatus]	62,897
PIL36359.1 transporter [Ganoderma sinense ZZ0214-1]	BAD98307.1 laccase3 [Trametes versicolor]	71,735
PIL36359.1 transporter [Ganoderma sinense ZZ0214-1]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	25,043

Fonte: O autor

Figura 21 – **Consulta 5:** Resultado gráfico com potenciais enzimas encontradas no *Pleurotus pulmonarius* PM_ss5.



Fonte: O autor

Tabela 10 – **Consulta 5:** Similaridade das enzimas encontradas no *Pleurotus pulmonarius* PM_ss5 com o banco de dados Plasticome.

Enzima consultada	Enzima com atividade comprovada	Similaridade (%)
KAF4569369.1 laccase, multicopper oxidase, benzenediol:oxygen oxidoreductase [Pleurotus pulmonarius]	AAR21094.1 laccase [Pleurotus ostreatus]	76,953
KAF4569369.1 laccase, multicopper oxidase, benzenediol:oxygen oxidoreductase [Pleurotus pulmonarius]	BAD98307.1 laccase3 [Trametes versicolor]	61.31
KAF4569369.1 laccase, multicopper oxidase, benzenediol:oxygen oxidoreductase [Pleurotus pulmonarius]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	27,508
KAF4569379.1 laccase [Pleurotus pulmonarius]	AAR21094.1 laccase [Pleurotus ostreatus]	59,841
KAF4569379.1 laccase [Pleurotus pulmonarius]	BAD98307.1 laccase3 [Trametes versicolor]	54.44
KAF4569379.1 laccase [Pleurotus pulmonarius]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	24,434
KAF4569591.1 laccase, multicopper oxidase, benzenediol:oxygen oxidoreductase [Pleurotus pulmonarius]	AAR21094.1 laccase [Pleurotus ostreatus]	60,588
KAF4569591.1 laccase, multicopper oxidase, benzenediol:oxygen oxidoreductase [Pleurotus pulmonarius]	BAD98307.1 laccase3 [Trametes versicolor]	55,556
KAF4569591.1 laccase, multicopper oxidase, benzenediol:oxygen oxidoreductase [Pleurotus pulmonarius]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	32.99
KAF4576943.1 laccase [Pleurotus pulmonarius]	AAR21094.1 laccase [Pleurotus ostreatus]	50,595
KAF4576943.1 laccase [Pleurotus pulmonarius]	BAD98307.1 laccase3 [Trametes versicolor]	46,139
KAF4576943.1 laccase [Pleurotus pulmonarius]	RMZ41798.1 multicopper oxidase/laccase [Aspergillus flavus]	23,364

Fonte: O autor

5 TRABALHOS FUTUROS

O Plasticome tem demonstrado sucesso na identificação de enzimas com potencial para degradação de plásticos em genomas fúngicos. No entanto, há áreas que podem ser otimizadas e expandidas para aprimorar ainda mais o desempenho da ferramenta.

Uma melhoria significativa seria a incorporação de uma funcionalidade durante o processo de BLAST que se concentre exclusivamente no sítio ativo das enzimas identificadas. Isso possibilitaria uma análise mais refinada, resultando em uma correspondência mais precisa entre as enzimas mineradas e aquelas já conhecidas.

A inclusão do SignalP no Plasticome representa um avanço promissor. Esta ferramenta é crucial para a pesquisa em fungos, permitindo a identificação de proteínas secretadas, especialmente relevantes para processos biológicos. No contexto do Plasticome, a integração do SignalP poderia aumentar a confiabilidade na identificação de enzimas potenciais, já que a capacidade de uma enzima degradar substâncias está diretamente ligada à sua secreção pelo fungo. Assim, a inclusão do SignalP enriqueceria a análise, proporcionando uma compreensão mais profunda das proteínas envolvidas nos processos de degradação de plásticos (TEUFEL et al., 2022).

Para ampliar a flexibilidade e utilidade do Plasticome, seria vantajoso permitir que os usuários enviassem seus próprios arquivos FASTA contendo sequências de proteínas. Atualmente restrito a organismos anotados no GenBank, essa expansão permitiria que pesquisadores explorassem uma variedade mais ampla de dados proteicos. Essa melhoria não apenas democratizaria o acesso à ferramenta, mas também enriqueceria a diversidade de dados analisados, ampliando assim o potencial para descobertas significativas.

Estas sugestões buscam não apenas otimizar tecnicamente o Plasticome, mas também expandir suas capacidades para atender às diversas demandas dos pesquisadores.

6 CONSIDERAÇÕES FINAIS

A pesquisa em torno do Plasticome representa uma etapa significativa no avanço da micorremediação por meio da identificação computacional de enzimas potencialmente capazes de degradar plásticos em genomas fúngicos. Os resultados obtidos, ao analisar diversos fungos, fornecem insights valiosos sobre o potencial generalizado desses organismos na luta contra a poluição plástica. Uma observação intrigante proveniente desta pesquisa é a identificação consistente de potenciais enzimas de degradação em todos os cinco fungos analisados. Essa constância levanta duas possibilidades, a primeira sugere que o método computacional adotado pode gerar falsos positivos, indicando a necessidade contínua de refinamento e validação das ferramentas utilizadas. A segunda possibilidade, aponta para a hipótese de que as enzimas com capacidade de degradar plásticos são mais comuns e difundidas entre os fungos do que se imaginava inicialmente. Essa descoberta provocativa sugere a necessidade de investigações mais aprofundadas, tanto experimentalmente quanto computacionalmente, para elucidar a prevalência real dessas enzimas na natureza. Diante dessas considerações, a pesquisa do Plasticome não apenas oferece uma ferramenta valiosa para a comunidade científica, mas também estabelece bases para futuras explorações no campo da micorremediação. O desafio agora reside em aprimorar as metodologias computacionais, validar os resultados experimentalmente e explorar ainda mais a diversidade fúngica em busca de soluções inovadoras para a problemática global dos resíduos plásticos.

REFERÊNCIAS

- BBC. A impressionante foto que mudou a percepção mundial sobre a crise do plástico. 2023. Disponível em: <<https://www.bbc.com/portuguese/articles/cl5z199ex1do>>.
- BINI, A.; OLIVEIRA, L.; POLISELO, H.; GUIMARÃES, F.; KASHIWABARA, A. Refatorando um pipeline de bioinformática: Um estudo de caso para análise de amplicons. In: . [S.l.: s.n.], 2021. p. 137–140.
- BIOPYTHON. 2023. <<https://biopython.org/>>. Accessed: 30/05/2023.
- CAZY. 2023. <<http://www.cazy.org/>>. Accessed: 30/05/2023.
- CELLERY. 2023. <<https://docs.celeryq.dev/en/stable/>>. Accessed: 30/05/2023.
- COCK, P. J. A.; ANTAO, T.; CHANG, J. T.; CHAPMAN, B. A.; COX, C. J.; DALKE, A.; FRIEDBERG, I. et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. **Bioinformatics**, Oxford University Press, v. 25, n. 11, p. 1422–1423, 2009. Disponível em: <<https://doi.org/10.1093/bioinformatics/btp163>>.
- CORRÊA, F. C. R. Mineração de dados como ferramenta para análise de base de dados de genoma do vírus influenza a. **Repositório Institucional da Universidade Federal de Ciências da Saúde de Porto Alegre**, 2017. Disponível em: <<https://repositorio.ufcspa.edu.br/jspui/handle/123456789/566>>.
- CRICK, F. Central dogma of molecular biology. **Nature Biotechnology**, 1970. Disponível em: <<https://www.nature.com/articles/227561a0#citeas>>.
- DALKIRAN, A.; RIFADIOGLU, A. S.; MARTIN, M. J.; CETIN-ATALAY, R.; ATALAY, V.; DOĞAN, T. Ecpred: a tool for the prediction of the enzymatic functions of protein sequences based on the ec nomenclature. **BMC Bioinformatics**, v. 19, n. 1, p. 334, 2018.
- DALY, P.; CAI, F.; KUBICEK, C. P.; JIANG, S.; GRUJIC, M.; RAHIMI, M. J.; SHE-TEIWI, M. S. et al. From lignocellulose to plastics: Knowledge transfer on the degradation approaches by fungi. **Biotechnology Advances**, Elsevier, v. 50, p. 107770, 2021.
- FLASK. 2023. <<https://flask.palletsprojects.com/en/3.0.x/>>. Accessed: 30/05/2023.
- FLURY, M.; NARAYAN, R. Biodegradable plastic as an integral part of the solution to plastic waste pollution of the environment. **Current Opinion in Green and Sustainable Chemistry**, v. 30, p. 100490, 2021. ISSN 2452-2236. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S2452223621000468>>.
- GALAL, A.; ELHASSAN, S. A.; SALEH, A. H.; AHMED, A. I.; ABDELRAHMAN, M. M.; KAMAL, M. M.; KHALEL, R. S.; ZIKO, L. A survey of the biosynthetic potential and specialized metabolites of archaea and understudied bacteria. **Current Research in Biotechnology**, Elsevier, v. 5, p. 100117, 2023.
- GENBANK. 2023. <<https://www.ncbi.nlm.nih.gov/genbank/>>. Accessed: 30/05/2023.

GOECKS, L. S.; SOUZA, M. d.; LIBRELATO, T. P.; TRENTTO, L. R. Design science research in practice: review of applications in industrial engineering. **Gestão Produção**, Universidade Federal de São Carlos, v. 28, n. 4, p. e5811, 2021. ISSN 0104-530X. Disponível em: <<https://doi.org/10.1590/1806-9649-2021v28e5811>>.

MARTIN, C.; FUSI, M.; TURNER, A.; ARIAS-ANDRES, M.; STIEGLITZ, T.; GODOY, V.; HOFMANN, T.; PINNA, M.; HOLMAN, I. Exponential increase of plastic burial in mangrove sediments as a major plastic sink. **Science Advances**, American Association for the Advancement of Science, v. 6, n. 23, p. eaaz5593, 2020. Disponível em: <<https://doi.org/10.1126/sciadv.aaz5593>>.

MCGINNIS, S.; MADDEN, T. L. Blast: at the core of a powerful and diverse set of sequence analysis tools. **Nucleic acids research**, Oxford University Press, v. 32, n. Web Server issue, p. W20–W25, 2004.

MIYAUCHI, S.; KISS, E.; KUO, A.; DRULA, E.; KOHLER, A.; SÁNCHEZ-GARCÍA, M.; MORIN, E.; ANDREOPOULOS, B.; BARRY, K. W.; BONITO, G. et al. Large-scale genome sequencing of mycorrhizal fungi provides insights into the early evolution of symbiotic traits. **Nature Communications**, Nature Publishing Group, v. 11, n. 1, p. 1–12, 2020.

MOHANTA, T. K.; BAE, H. The diversity of fungal genome. **Biological Procedures Online**, v. 17, n. 1, p. 8, 2015. Disponível em: <<https://doi.org/10.1186/s12575-015-0020-z>>.

MORGAN, T. **ATA DA DEFESA DE TESE**. 2021. Acesso em: <https://repositorio.ufmg.br/bitstream/1843/36634/1/tese_final_tulio_morgan.pdf>.

NCBI blast tutorial. 2023. <<https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html#head3>>. Accessed: 23/11/2023.

PYTHON. 2023. <<https://www.python.org/>>. Accessed: 30/05/2023.

RABBITMQ. 2023. <<https://www.rabbitmq.com/>>. Accessed: 30/05/2023.

SHELDRAKE, M. **Entangled Life: How Fungi Make Our Worlds, Change Our Minds & Shape Our Futures**. [S.l.]: Random House, 2020.

TEMPORITI, M. E. E.; NICOLA, L.; NIELSEN, E.; TOSI, S. Fungal enzymes involved in plastics biodegradation. **Microorganisms**, v. 10, n. 6, 2022. ISSN 2076-2607. Disponível em: <<https://www.mdpi.com/2076-2607/10/6/1180>>.

TEUFEL, F.; ARMENTEROS, J. J. A.; JOHANSEN, A. R.; GÍSLASON, M. H.; PIHL, S. I.; TSIRIGOS, K. D.; WINTHER, O.; BRUNAK, S.; HEIJNE, G. von; NIELSEN, H. Signalp 6.0 predicts all five types of signal peptides using protein language models. **Nature Biotechnology**, v. 40, p. 1023–1025, 2022. Disponível em: <<https://www.nature.com/articles/s41587-021-01156-3>>.

THUSHARI, G. G. N.; SENEVIRATHNA, J. D. M. Plastic pollution in the marine environment. **Heliyon**, Elsevier, v. 6, n. 8, p. e04709, 2020.

UNIPROT. 2023. <<https://www.uniprot.org/>>. Accessed: 30/05/2023.

WISESO, L. G.; IMRONA, M.; ALAMSYAH, A. Performance analysis of neo4j, mongodb, and postgresql on 2019 national election big data management database. **2020 6th International Conference on Science in Information Technology (ICSITech)**, p. 91–96, 2020.

YADAV, P.; RAI, S. N.; MISHRA, V.; GUPTA, A.; KAUSHIK, P. Mycoremediation of environmental pollutants: a review with special emphasis on mushrooms. **Environmental Sustainability**, Springer, v. 4, n. 4, p. 605–618, 2021.

ZHANG, C.; LIU, B.-y.; LIU, J.-w.; YAN, D.-j.; BAI, J.; ZHANG, Y.-l.; MOU, Y.-h.; HU, Y.-c. Gene mining and efficient biosynthesis of a fungal peptidyl alkaloid. **Chinese Herbal Medicines**, v. 13, n. 1, p. 98–104, Jan 2021.

ZHENG, J.; GE, Q.; YAN, Y.; ZHANG, X.; HUANG, L.; YIN, Y. dbcan3: automated carbohydrate-active enzyme and substrate annotation. **Nucleic Acids Research**, 2023. Disponível em: <<https://doi.org/10.1093/nar/gkad328>>.

ZHOU, L.; SONG, C.; LI, Z.; KUIPERS, O. P. Antimicrobial activity screening of rhizosphere soil bacteria from tomato and genome-based analysis of their antimicrobial biosynthetic potential. **BMC genomics**, BioMed Central, v. 22, n. 1, p. 29, 2021.