



UNIVERSIDADE DO ESTADO DA BAHIA (UNEB)
DEPARTAMENTO DE CIÊNCIAS EXATAS E DA TERRA, CAMPUS I
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

RENILSON BISPO DE ASSIS SOUZA

**KINGFUNGI: UM PIPELINE PARA PROSPEÇÃO DE COGUMELOS
COM CAPACIDADE DE ACÚMULO DE MICRONUTRIENTES**

SALVADOR
2024

RENILSON BISPO DE ASSIS SOUZA

**KINGFUNGI: UM PIPELINE PARA PROSPECÇÃO DE COGUMELOS
COM CAPACIDADE DE ACÚMULO DE MICRONUTRIENTES**

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação do Departamento de Ciências Exatas e da Terra (DCET) - Campus I, da Universidade do Estado da Bahia (UNEBA), como requisito à obtenção do grau de bacharel em Sistemas de Informação.

Área de concentração: Bioinformática

Orientador: Prof. Dr. Alexandre Rafael Lenz

SALVADOR
2024

REINILSON BISPO DE ASSIS SOUZA

**KINGFUNGI: UM PIPELINE PARA PROSPECÇÃO DE COGUMELOS COM
CAPACIDADE DE ACÚMULO DE MICRONUTRIENTES**

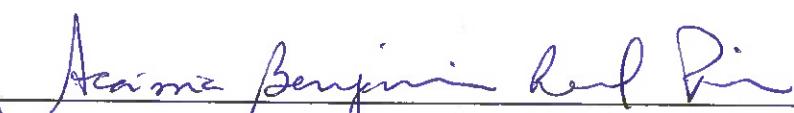
Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação do Departamento de Ciências Exatas e da Terra (DCET) - Campus I, da Universidade do Estado da Bahia (UNEBA), como requisito à obtenção do grau de bacharel em Sistemas de Informação.
Área de concentração: Bioinformática

Aprovada em: .

BANCA EXAMINADORA


Prof. Dr. Alexandre Rafael Lenz
Orientador


Prof. Dr. Vagner de Souza Fonseca
Examinador interno (DCET-I/UNEBA)


Prof. Dra. Acássia Benjamin Leal Pires
Examinador interno (DCV-I/UNEBA)

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus por ter me dado forças para continuar, mesmo nos dias em que elas estavam quase no fim. À minha mãe, Sra. Elizangela Bispo de Assis Souza, pelo amor incondicional, pelo apoio constante e pelos muitos sacrifícios feitos ao longo da vida para que eu pudesse chegar até aqui, minha eterna gratidão. Aos meus irmãos Remisson Bispo, Rosemire Bispo e Ismael Bispo, meu sincero agradecimento. Vocês são minha base sólida e sempre me encorajaram a persistir, mesmo nos momentos mais desafiadores. Sou imensamente grato por tê-los ao meu lado.

Ao meu orientador, Professor Dr. Alexandre Rafael Lenz, um exemplo de pessoa, expresso minha gratidão pela paciência, amizade, orientação e conselhos transmitidos ao longo deste trabalho. Seu comprometimento e dedicação foram essenciais para o desenvolvimento deste projeto. Aproveito o momento para agradecer grandemente à professora Maria Cristina Elyote Santos, que acompanhou este trabalho na primeira etapa, na disciplina de TCC1, e à professora Débora Alcina Rego Chaves que, com toda paciência, carinho e respeito, me ajudou neste trabalho.

Aos meus colegas de classe e amigos, especialmente Vitor Manoel, pelos vários dias me ouvindo dizer que iria desistir e voltar para casa, e por sempre me fazer desistir de desistir. Jéssica Cardoso e Filipe Silva, pelos vários dias em que passávamos 10, 13 e até 17 horas fazendo os trabalhos, agradeço pelo incentivo contínuo e pela troca de experiências. Vocês, com certeza, fizeram toda a diferença. Agradeço também por todas as discussões enriquecedoras que tivemos.

" [...] É que um neguinho igual eu tem que ser duas vezes mais astuto, mais dedicado no estudo, mais focado que eles tudo, é que a porra desse mundo odeia o nosso rosto mulato."

(Leal Mc)

RESUMO

Atualmente, no mundo, o problema da fome assombra milhares de lares. No Brasil, no ano de 2022, foi registrado que 33,1 milhões de brasileiros viviam em insegurança alimentar. Cogumelos comestíveis podem revelar uma possível solução para esse problema através de novas fontes de alimento, levando em consideração que podem ser produzidos em larga escala em ambiente urbano e com poucos recursos. Além disso, os cogumelos podem acumular micronutrientes e outros elementos essenciais para nutrição humana. Este estudo descreveu o desenvolvimento da aplicação web *KINGFUNGI*, visando à identificação de proteínas transportadoras de micronutrientes e outros elementos essenciais à saúde humana em genomas de fungos, especialmente cogumelos. Com base em uma revisão abrangente da literatura, foram empregadas técnicas avançadas de bioinformática e análise de dados genômicos para propor um método computacional destinado à localização de genes-alvo nos genomas fúngicos. Adotando a abordagem de *Design Science Research (DSR)*, essa metodologia gerou um artefato tecnológico (*KingFungi*), sustentado por Objetivos e Resultados Chave (OKR) para monitorar o progresso do projeto através da plataforma Trello. O *KINGFUNGI* desdobrou-se em seis estágios interconectados, utilizando o módulo *Chain* do *Celery* para análise sequencial e ordenada de genomas de fungos. Para isso, foram utilizados os softwares *DeepTMHMM*, para identificar proteínas transmembrana, e *Mebipred*, para identificar proteínas que se ligam a íons de metal. A validação do método proposto, utilizando dados do *GenBank*, empregou 3 genomas de fungos do filo *Basidiomycota*, espécies de cogumelos comestíveis de 3 diferentes ordens, demonstrando sua eficácia e precisão na identificação de proteínas associadas à absorção e acumulação de micronutrientes e outros elementos essenciais em cogumelos através da ligação a íons de metal. As análises realizadas constataram que o *KINGFUNGI* conseguiu identificar 10 elementos, sendo 8 micronutrientes e 2 elementos essenciais. Como exemplo, temos o proteoma do cogumelo *Agaricus bisporus*, popularmente conhecido como Cogumelo Paris ou *Champignon*, que revelou a presença de 1536 proteínas transmembrana. Destas, foram identificadas apenas aquelas que têm capacidade de se ligar a íons metálicos, com as seguintes quantidades previstas: cálcio (5 proteínas), cobalto (6 proteínas), cobre (20 proteínas), ferro (9 proteínas), potássio (30 proteínas), magnésio (2 proteínas), sódio (1 proteína) e zinco (4 proteínas) entre os micronutrientes, e manganês (20 proteínas) e níquel (30 proteínas) entre os elementos essenciais. Esta pesquisa pode contribuir de forma significativa para o avanço do conhecimento sobre espécies de cogumelos comestíveis e como a suplementação do substrato para produção pode gerar novas fontes de alimento ricas em nutrientes, podendo contribuir para a mitigação da insegurança alimentar. O *KingFungi* está disponível no repositório do Grupo de Pesquisa em Bioinformática e Biologia Computacional- G2BC, GitHub: <https://github.com/G2BC/KingFungi>.

Palavras-chave: Inteligência Artificial; *Deep Learning*; Genoma; Genes; Mineração em genoma; Micronutrientes.

ABSTRACT

Currently, around the world, the problem of hunger haunts thousands of homes. In Brazil, in 2022, it was recorded that 33.1 million Brazilians lived in food insecurity. Edible mushrooms may reveal a possible solution to this problem through new food sources, taking into account that they can be produced on a large scale in urban environments with few resources. Furthermore, mushrooms can accumulate micronutrients and other elements essential for human nutrition. This study described the development of the *KINGFUNGI* web application, aimed at identifying proteins transporting micronutrients and other elements essential to human health in fungal genomes, especially mushrooms. Based on a comprehensive literature review, advanced bioinformatics and genomic data analysis techniques were employed to propose a computational method aimed at localizing target genes in fungal genomes. Adopting the *Design Science Research (DSR)* approach, this methodology generated a technological artifact (*KingFungi*), supported by Objectives and Key Results (OKR) to monitor the project's progress through the Trello platform. *KINGFUNGI* unfolded into six interconnected stages, using the *Chain* module of *Celery* for sequential and ordered analysis of fungal genomes. For this, the software *DeepTMHMM* was used, to identify transmembrane proteins, and *Mebipred*, to identify proteins that bind metal ions. Validation of the proposed method, using data from *GenBank*, used 3 genomes of fungi from the phylum *Basidiomycota*, species of edible mushrooms from 3 different orders, demonstrating its effectiveness and precision in identifying proteins associated with absorption and accumulation of micronutrients and other essential elements in mushrooms through binding to metal ions. The analyzes carried out found that *KINGFUNGI* managed to identify 10 elements, 8 micronutrients and 2 essential elements. As an example, we have the proteome of the mushroom *Agaricus Bisporus*, popularly known as Paris Mushroom or *Champignon*, which revealed the presence of 1536 transmembrane proteins. Of these, only those that have the capacity to bind to metallic ions were identified, with the following predicted quantities: calcium (5 proteins), cobalt (6 proteins), copper (20 proteins), iron (9 proteins), potassium (30 proteins), magnesium (2 proteins), sodium (1 protein) and zinc (4 proteins) among the micronutrients, and manganese (20 proteins) and nickel (30 proteins) among the essential elements. This research can significantly contribute to advancing knowledge about edible mushroom species and how supplementing the substrate for production can generate new nutrient-rich food sources, which can contribute to mitigating food insecurity. *KingFungi* is available in the repository of the Bioinformatics and Computational Biology Research Group - G2BC, GitHub: <https://github.com/G2BC/KingFungi>.

Key-words: Artificial Intelligence; Deep Learning; Genome; Genes; Genome Mining; Micronutrients.

LISTA DE ILUSTRAÇÕES

| | |
|---|----|
| Figura 1 – Cogumelos comestíveis | 20 |
| Figura 2 – Dogma Central da Biologia Molecular | 26 |
| Figura 3 – Processo de Transporte | 29 |
| Figura 4 – Processo de ligação de proteínas | 30 |
| Figura 5 – Redes neural convolucional. | 35 |
| Figura 6 – Redes Neurais Recorrentes | 36 |
| Figura 7 – Deep Residual Network. | 36 |
| Figura 8 – Arquitetura do DeepTMHMM | 52 |
| Figura 9 – Exemplo do processamento do DeepTMHMM | 53 |
| Figura 10 – <i>Design Science Research (DSR)</i> | 55 |
| Figura 11 – Quadro Kanban. | 56 |
| Figura 12 – Interface do Trello. | 57 |
| Figura 13 – Representação gráfica das etapas do pipeline | 58 |
| Figura 14 – Front-end KingFungi Mobile | 59 |
| Figura 15 – Front-end KingFungi | 60 |
| Figura 16 – Arquitetura do sofware | 61 |
| Figura 17 – Estrutura de um arquivo .fasta | 64 |
| Figura 18 – Exemplo do arquivo de saída do DeepTMHMM(.3line) | 65 |
| Figura 19 – Exemplo do arquivo de saída do analisador DeepTMHMM(deep_out.fasta) | 65 |
| Figura 20 – Exemplo do arquivo de saída do analisador Mymetal | 67 |
| Figura 21 – E-mail de resultados | 68 |
| Figura 22 – Gráfico do <i>Agaricales Agaricus bisporus</i> | 70 |
| Figura 23 – Resultados da análise no arquivo TSV do <i>Agaricales Agaricus bisporus</i> | 71 |
| Figura 24 – Gráfico do <i>Auriculariales Exidia glandulosa</i> | 71 |
| Figura 25 – Resultados da análise no arquivo TSV do <i>Auriculariales Exidia glandulosa</i> | 72 |
| Figura 26 – Gráfico do <i>Boletales Boletus edulis</i> | 72 |
| Figura 27 – Resultados da análise no arquivo TSV do <i>Boletales Boletus edulis</i> | 73 |
| Figura 28 – Tempo de execução da análise dos 3 genomas | 73 |
| Figura 29 – Tempo de execução do MebiPred de forma isolada | 74 |
| Figura 30 – Representação gráfica das etapas do pipeline pós remodelagem | 75 |
| Figura 31 – Tempo de execução da análise dos 3 genomas pós remodelagem | 76 |

LISTA DE TABELAS

| | |
|---|----|
| Tabela 1 – Palavras-chave e sinônimos utilizados nas buscas | 42 |
| Tabela 2 – Critérios de inclusão e exclusão dos artigos | 43 |
| Tabela 3 – Tabela de Ferramentas e aplicações | 43 |
| Tabela 4 – Organismos selecionados para a avaliação do KingFungi (KF) | 69 |

LISTA DE QUADROS

| | |
|--|----|
| Quadro 1 – Micronutrientes e elementos essenciais | 24 |
| Quadro 2 – Acúmulo de elementos inorgânicos em cogumelos e predição de proteínas que se ligam a íons metálicos | 50 |

SUMÁRIO

| | | |
|---------|---|----|
| 1 | INTRODUÇÃO | 13 |
| 2 | REFERENCIAL TEÓRICO | 17 |
| 2.1 | REFERENCIAL DA ÁREA DE APLICAÇÃO | 17 |
| 2.1.1 | <i>Insegurança alimentar (Fome)</i> | 17 |
| 2.1.2 | <i>Cogumelos comestíveis</i> | 18 |
| 2.1.3 | <i>Micronutrientes e outros elementos essenciais para nutrição humana</i> | 22 |
| 2.1.3.1 | <i>Cogumelos como fonte de micronutrientes e outros elementos essenciais para nutrição humana</i> | 25 |
| 2.1.4 | <i>Genômica de Fungos</i> | 25 |
| 2.1.4.1 | <i>Proteínas transportadoras</i> | 28 |
| 2.1.4.2 | <i>Proteínas que se ligam a íons de metais</i> | 29 |
| 2.2 | REFERENCIAL COMPUTACIONAL | 30 |
| 2.2.1 | <i>Bioinformática</i> | 30 |
| 2.2.1.1 | <i>Pipeline de Bioinformática</i> | 31 |
| 2.2.1.2 | <i>Mineração em genomas</i> | 32 |
| 2.2.2 | <i>Redes Neurais Artificiais</i> | 33 |
| 2.2.2.1 | <i>Modelos de Redes Neurais</i> | 35 |
| 2.2.2.2 | <i>Aprendizado Profundo (Deep Learning)</i> | 36 |
| 2.2.3 | <i>Aplicações Web para Bioinformática</i> | 37 |
| 2.2.3.1 | <i>Python e Biopython</i> | 38 |
| 2.2.3.2 | <i>Genbank</i> | 39 |
| 2.2.3.3 | <i>Celery</i> | 40 |
| 2.2.3.4 | <i>RabbitMQ</i> | 40 |
| 3 | TRABALHOS CORRELATOS | 42 |
| 3.1 | Estado da Arte | 42 |
| 3.1.1 | <i>Mebipred (Mymetal)</i> | 48 |
| 3.1.2 | <i>DeepTMHMM</i> | 50 |
| 4 | METODOLOGIA | 54 |
| 4.1 | Metodologia científica | 54 |
| 4.2 | Metodologia de Desenvolvimento | 55 |
| 5 | RESULTADOS | 58 |
| 5.1 | Método Computacional(Pipeline) | 58 |
| 5.1.1 | <i>Arquitetura da aplicação</i> | 58 |
| 5.1.1.1 | <i>Back-end</i> | 61 |
| 5.1.1.2 | <i>Front-end</i> | 62 |

| | | |
|-------|---|----|
| 5.1.2 | <i>Workflow de execução da aplicação</i> | 63 |
| 5.2 | Avaliação de eficiência (tempo de execução) e eficácia (funcionalidade) | 68 |
| 5.2.1 | <i>Ambiente de avaliação e genomas utilizados</i> | 68 |
| 5.2.2 | <i>Avaliação 1 (Funcionalidades)</i> | 70 |
| 5.2.3 | <i>Avaliação 2 (Tempo)</i> | 73 |
| 5.3 | Remodelagem do pipeline | 74 |
| 5.3.1 | <i>Avaliação 1 (Funcionalidades) pós remodelagem</i> | 75 |
| 5.3.2 | <i>Avaliação 2 (Tempo) pós remodelagem</i> | 75 |
| 6 | TRABALHOS FUTUROS | 77 |
| 7 | CONSIDERAÇÕES FINAIS | 78 |
| | REFERÊNCIAS | 80 |

1 INTRODUÇÃO

A tecnologia de sequenciamento de nova geração *Next Generation Sequencing* (NGS), revolucionou o campo da genômica, permitindo análises rápidas, eficientes e em larga escala de grandes quantidades de informações genéticas (Vincent *et al.*, 2017). Essa abordagem tem sido fundamental para avanços significativos em áreas como medicina personalizada, diagnóstico de doenças genéticas, estudos de populações e evolução, biologia do câncer, entre outras. No entanto, o NGS também apresenta desafios, como a interpretação e extração de informações relevantes dos conjuntos de dados genômicos complexos.

Para lidar com esses desafios, a mineração de dados genômicos surgiu como uma área tecnológica que visa explorar e analisar as enormes quantidades de dados existentes (Rekadwad; Gonzalez, 2017). A aplicação da Inteligência Artificial (IA) e do Aprendizado Profundo na análise de dados genômicos tem desempenhado um papel fundamental no avanço desse campo, permitindo a identificação de padrões complexos e a realização de tarefas como classificação e interpretação dos dados genômicos. Essas técnicas avançadas têm proporcionado uma compreensão mais profunda da estrutura e função do genoma humano, abrindo novas oportunidades para avanços científicos e médicos (D'Agaro, 2018).

No contexto da mineração em genomas fúngicos, a análise de dados genômicos e o uso de técnicas de Aprendizado Profundo têm revelado informações valiosas. Por exemplo, (Baltz, 2021) utiliza essa técnica para a descoberta de medicamentos. Além disso, podemos obter conhecimento sobre a estrutura genética dos cogumelos e suas potenciais capacidades. Alguns cogumelos apresentam a capacidade única de absorver e acumular micronutrientes essenciais em seus tecidos, o que é de grande interesse para a nutrição humana e a saúde (Jiménez *et al.*, 2024; Beelman; Kalaras; Jr, 2019). Esse potencial levanta questões importantes sobre a adaptabilidade dos cogumelos a diferentes ambientes e seu papel na promoção da saúde humana através da ingestão de alimentos (Beelman; Kalaras; Jr, 2019).

Portanto, a identificação dos genes envolvidos no processo de transporte e acumulação de micronutrientes em genomas fúngicos pode permitir uma compreensão mais aprofundada dos mecanismos moleculares subjacentes a essa capacidade singular dos cogumelos, que tenta resolver o problema de pesquisa encontrado e descrito como: “Como desenvolver um pipeline para prospecção de cogumelos com capacidade de acúmulo de micronutrientes ?”.

O objetivo geral deste trabalho é desenvolver um método computacional, baseado em IA, capaz de minerar genomas de cogumelos para encontrar proteínas transportadoras de micronutrientes e elementos essenciais para nutrição humana. Para alcançar esse objetivo

geral, foram estabelecidos os seguintes objetivos específicos:

- Propor um método computacional para a identificação dos genes alvo nos genomas de cogumelos;
- Avaliar a eficiência do método em termos de tempo de execução para genomas de cogumelos de diferentes ordens.

A utilização de técnicas avançadas de IA e Aprendizado Profundo é fundamental para a eficiente análise de grandes conjuntos de dados genômicos. Por meio dessas técnicas, é possível identificar padrões genéticos (Anguita-Ruiz *et al.*, 2020) relevantes relacionados ao transporte e ao acúmulo de micronutrientes, impulsionando assim o campo da genômica fúngica. Esses *insights* valiosos poderão ser aplicados em diversos setores, como nutrição, saúde pública, biotecnologia e agricultura.

Para atingir os objetivos propostos, este projeto utilizou DSR como metodologia para solucionar o problema existente. Inicialmente, foi necessária a realização de uma revisão sistemática da literatura, com o intuito de compreender o estado da arte sobre o tema em questão. Essa revisão permitiu uma assimilação apropriada do conhecimento existente e a identificação de lacunas a serem preenchidas.

Para solucionar o problema existente, foi proposto e desenvolvido um pipeline computacional para a mineração em genomas de fungos, levando em consideração os padrões genéticos relevantes associados ao transporte de micronutrientes para dentro da célula através das proteínas transportadoras de metais.

Para validar o método proposto, foram utilizados dados provenientes do banco de dados genômico GenBank, que armazena sequências genômicas. Através dessa validação, foi possível verificar a eficiência e a eficácia do método desenvolvido e alcançar os objetivos deste trabalho.

Diante do exposto, o projeto vai otimizar o processo de identificação da capacidade de acumular micronutrientes em fungos do filo *Basidiomycota*, especialmente em cogumelos comestíveis. A prospecção computacional de espécies de cogumelos que acumulam micronutrientes importantes para nutrição humana contribuirá para a ampliação do conhecimento sobre essas espécies, permitindo a identificação dos genes responsáveis pela acumulação de elementos inorgânicos. Os cogumelos são matriz alternativas de grande interesse medicinal e nutricional, e a identificação dos genes relacionados a essas características oferecerá oportunidades para o desenvolvimento de aplicações terapêuticas e nutricionais inovadoras (Reis *et al.*, 2017; Barbosa *et al.*, 2020; Venturella *et al.*, 2021).

Além disso, o avanço na genômica fúngica, impulsionado por esse projeto, levará a comunidade científica a entender ainda mais o campo da biotecnologia e auxiliará na

identificação de fungos capazes de acumular micronutrientes, fornecendo informações importantes para a resolução ou diminuição de problemas sociais existentes.

No setor industrial, a compreensão dos mecanismos moleculares subjacentes ao transporte de micronutrientes em genomas fúngicos poderá levar ao desenvolvimento de novas estratégias de produção e processamento de alimentos ricos em micronutrientes, contribuindo para a eficiência e sustentabilidade desses processos. Dessa forma, este projeto representa uma oportunidade promissora para a integração da genômica fúngica, IA e Aprendizado Profundo, com benefícios abrangentes para a ciência, tecnologia, economia, sociedade e meio ambiente.

Além dos interesses medicinal e biotecnológico, destaca-se o valor nutricional dos cogumelos. Por exemplo, o estudo realizado por (Bito *et al.*, 2014) revelou que os cogumelos comestíveis consumidos por vegetarianos europeus, como *Craterellus cornucopoides* e *Cantharellus cibarius*, contêm níveis consideráveis de vitamina B12. Além disso, foi demonstrado que os *Lentinula edodes*, conhecidos popularmente como cogumelos shiitake, também possuem teores significativos de vitamina B12. A qualidade das proteínas dos cogumelos é superior às proteínas vegetais (Cheung, 2010). Cogumelos são particularmente notáveis por seu baixo teor calórico e alto conteúdo de fibras dietéticas, o que os torna um complemento ideal para dietas saudáveis e balanceadas (Kumari, 2020; Cheung, 2010). A inclusão de cogumelos na alimentação pode contribuir significativamente para a ingestão de nutrientes essenciais, promovendo benefícios à saúde como fortalecimento do sistema imunológico, propriedades anti-inflamatórias e potencial efeito anticancerígeno (Rizzo *et al.*, 2021).

Portanto, os cogumelos comestíveis podem desempenhar um papel importante na mitigação do problema da insegurança alimentar. O Objetivo de Desenvolvimento Sustentável 2 (ODS 2) da ONU, que visa erradicar a fome, alcançar a segurança alimentar e melhorar a nutrição, destaca a importância de alimentos nutritivos como os cogumelos (Kumari, 2020; Cheung, 2010). Igualmente importante, o cultivo de cogumelos é uma prática agrícola sustentável que pode ser realizada em pequenas áreas, utilizando resíduos agrícolas como substrato (Jones, 2021). Isso não só gera alimentos de alto valor nutritivo, como também promove a melhoria da qualidade alimentar. Ademais, a produção de cogumelos pode ser uma fonte de renda para pequenos produtores, inclusive em regiões urbanas, auxiliando no combate à pobreza e melhorando a segurança alimentar (Jones, 2021).

A motivação para o desenvolvimento deste projeto reside na aplicação dos conhecimentos adquiridos durante a formação universitária para a criação de soluções computacionais que auxiliem na resolução de problemas sociais. Um dos desafios a serem enfrentados é a promoção da saúde humana através da ingestão de alimentos ricos em micronutrientes, contribuindo para o bem-estar e qualidade de vida da população.

Pesquisas demonstram que os cogumelos comestíveis podem contribuir para a melhoria da qualidade alimentar, sendo crucial compreender o impacto positivo desses micronutrientes na saúde humana para a identificação de estratégias eficazes de promoção da saúde e prevenção de doenças.

Os resultados esperados para esta pesquisa são:

- Revisão sistemática da literatura;
- Aplicação web implementando o método computacional proposto;
- Avaliação de eficiência (tempo de execução) e eficácia (funcionalidade) da aplicação *web*.

Ao atingir esses objetivos, espera-se contribuir para o avanço do conhecimento científico na área da bioinformática.

O presente trabalho é organizado de forma a proporcionar uma compreensão clara e progressiva dos conceitos e resultados obtidos ao longo da pesquisa. A seguir, é apresentada uma visão geral da estrutura dos capítulos que compõem este estudo: No Capítulo 1, a Introdução, são apresentados os conceitos iniciais da pesquisa, incluindo o tema, a motivação, os objetivos e a justificativa do estudo. Esta seção estabelece o contexto e a relevância do trabalho, oferecendo uma visão geral do problema abordado. O Capítulo 2, Referencial Teórico, é dividido em duas seções principais: Referencial da Área de Aplicação, que apresenta o referencial biológico necessário para a pesquisa, e Referencial Computacional, que abrange os conceitos, tecnologias e metodologias computacionais relevantes. O Capítulo 3 é dedicado aos Trabalhos Correlatos, onde são revisadas as ferramentas e estudos encontrados na literatura que têm relação com a pesquisa realizada. No Capítulo 4, Metodologia, é descrita a metodologia utilizada no desenvolvimento deste trabalho, detalhando os métodos e procedimentos adotados. O Capítulo 5 apresenta os Resultados da Pesquisa, incluindo as ferramentas utilizadas, o ambiente de desenvolvimento, a arquitetura do sistema e a avaliação do método computacional implementado. No Capítulo 6, são discutidas as Possibilidades de Trabalhos Futuros, explorando as potenciais extensões e melhorias que podem ser realizadas a partir dos achados desta pesquisa. Finalmente, o Capítulo 7 contém as Considerações Finais, onde são resumidas as conclusões do estudo e avaliadas as contribuições e implicações dos resultados obtidos.

2 REFERENCIAL TEÓRICO

Neste capítulo, serão apresentados os conceitos fundamentais do referencial teórico, organizados em duas seções distintas. A primeira seção abordará o referencial da área de aplicação da pesquisa, enquanto a segunda discorrerá sobre o referencial relacionado à parte computacional.

2.1 REFERENCIAL DA ÁREA DE APLICAÇÃO

Nesta seção serão apresentados conceitos do referencial teórico da área de aplicação da presente pesquisa.

2.1.1 *Insegurança alimentar (Fome)*

A insegurança alimentar emerge como uma problemática global de magnitude significativa, afetando milhões de pessoas em diversas partes do mundo. Este fenômeno complexo transcende fronteiras geográficas e socioeconômicas, manifestando-se de maneiras variadas e impactando diretamente a qualidade de vida das populações vulneráveis (Bezerra; Olinda; Pedraza, 2017).

No cerne da insegurança alimentar encontra-se a inadequação entre a disponibilidade de alimentos e as necessidades nutricionais da população. Fatores multifacetados contribuem para esse desequilíbrio, incluindo questões econômicas, políticas ambientais e sociais. Segundo Uauy (2005), as necessidades nutricionais dos seres humanos evoluíram ao longo do tempo, devido às mudanças nos padrões alimentares. A compreensão aprofundada desses elementos é crucial para a formulação de estratégias eficazes de combate à insegurança alimentar e importante para compreender o processo de envelhecimento humano (Maciel *et al.*, 2015).

A globalização econômica, por exemplo, desencadeia cadeias de produção e distribuição complexas, que nem sempre beneficiam as comunidades mais carentes. A disparidade de acesso a recursos e oportunidades amplifica as desigualdades, resultando em dificuldades crescentes para certas populações em garantir uma nutrição adequada para seus membros (Stoian *et al.*, 2016).

Além disso, eventos climáticos extremos e mudanças climáticas têm impacto direto na produção agrícola, exacerbando a insegurança alimentar. A volatilidade climática, associada à exploração inadequada dos recursos naturais, desafia a estabilidade dos sistemas alimentares, deixando as comunidades vulneráveis à escassez e aos preços elevados dos alimentos (Gulati *et al.*, 2009).

A dimensão política da insegurança alimentar também não pode ser negligenciada. Conflitos armados, instabilidade governamental e corrupção comprometem a capacidade dos Estados de fornecerem infraestrutura e serviços básicos, incluindo sistemas alimentares seguros e acessíveis. A falta de governança eficaz cria um ambiente propício para o agravamento da insegurança alimentar.

A insegurança alimentar não se limita apenas à privação física de alimentos, mas também abrange a insuficiência de acesso à dietas equilibradas e nutritivas. A desnutrição e a má alimentação resultam em consequências de longo prazo para a saúde das populações afetadas, perpetuando o ciclo de pobreza e vulnerabilidade. No Brasil, por exemplo, segundo dados de Silva *et al.* (2020), em uma pesquisa publicada em 2020, com dados retirados do IBGE (2019), o Brasil conta com 7 milhões de pessoas que vivem assombradas diariamente pela fome e que cerca de 40 milhões não possuem alimentação nutricionalmente adequada, o que configura a insegurança alimentar leve e em dados mais recentes (2022) pela Rede Brasileira de Pesquisa em Soberania e Segurança Alimentar e Nutricional (Rede Penssan), revelou que cerca de 58,7% da população, equivalente a 33,1 milhões de pessoas, estava em algum estágio de insegurança alimentar (Rede PENSSAN, 2022).

Diante desse panorama desafiador, é importante adotar uma abordagem integrada para enfrentar a insegurança alimentar. Estratégias que abordem questões econômicas, promovam práticas agrícolas sustentáveis, fortaleçam instituições políticas e fomentem a educação nutricional são essenciais para criar mudanças duradouras.

Além disso, a cooperação internacional desempenha um papel crucial na mitigação da insegurança alimentar. Parcerias entre ciência, governos, organizações não governamentais e setor privado são fundamentais para implementar soluções abrangentes e sustentáveis.

2.1.2 Cogumelos comestíveis

Os fungos formam um reino de organismos eucarióticos com ampla variação de aspectos morfológicos, moleculares, bioquímicos e ecológicos. Os fungos são uma das cinco principais linhagens multicelulares na árvore da vida, sendo que animais e fungos estão mais intimamente relacionados entre si do que os animais estão com as plantas terrestres. Eles são organismos fundamentais dos ecossistemas e têm recebido menos atenção do que a Flora e Fauna, embora onipresentes e de natureza altamente diversificada, formando seu próprio conjunto de organismos, a Funga (Kuhar *et al.*, 2018).

Cerca de 150.000 espécies de fungos foram descritas pela ciência (Hyde, 2022), mas a biodiversidade global do reino dos fungos está longe de ser totalmente compreendida. Uma estimativa de 2017 sugere que podem existir entre 2,2 e 3,8 milhões de espécies (Hawksworth; Lücking, 2017). O número de novas espécies de fungos descobertas vem aumentando anualmente. No ano de 2019, foram descritas 1.882 novas espécies de fungos e

foi estimado que mais de 90% dos fungos permanecem desconhecidos (Cheek *et al.*, 2020).

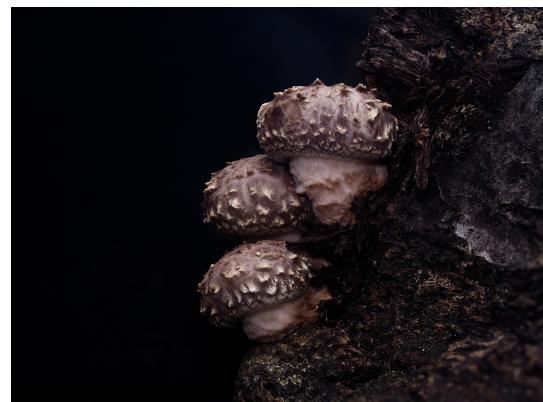
Algumas espécies dos filos Basidiomycota e Ascomycota desenvolvem estruturas reprodutivas macroscópicas, também conhecidas como macrofungos ou popularmente chamadas de cogumelos e orelhas-de-pau. Hawksworth e Lücking (2017) estimaram a existência de cerca de 22 mil espécies de cogumelos, sendo que, destas, aproximadamente 2.189 espécies são consideradas comestíveis, das quais 2.006 podem ser consumidas com segurança, e outras 183 espécies exigem alguma forma de pré-tratamento antes do consumo seguro ou foram associadas a reações alérgicas (Li *et al.*, 2021). Cogumelos comestíveis são coletados na natureza, como demonstrado na Figura 1, ou cultivados em todo o mundo (Thu *et al.*, 2020a). Pouco mais de 90 espécies comestíveis são cultivadas em todo o mundo, das quais cerca de 30 espécies são cultivadas comercialmente para alimentação e amplamente consumidas, enquanto que algumas poucas espécies são cultivadas para fins medicinais.

Os primeiros relatos de consumo de cogumelos são da Espanha (18.700 anos atrás), China (5.000 a 6.000 anos atrás) e Egito (4.600 anos atrás) (Li *et al.*, 2021). Os cogumelos têm uma longa história de usos por suas propriedades medicinais e nutricionais. Eles vêm sendo consumidos por humanos há milhares de anos e utilizados de forma medicinal há mais de 2.000 anos (Bell *et al.*, 2022).

Etnologicamente, a relação dos fungos com a população humana faz com que os povos, de maneira geral, sejam classificados como micófilos e não micófilos, ou seja, se eles têm interesse ou não por fungos. No Brasil, até a década de 60 do século passado, as populações indígenas eram consideradas não micófilas. A partir de então, algumas publicações revelaram o conhecimento etnomicológico de etnias da região amazônica, incluindo os Yanomami, Tucano, Nambiquara, Caibi, Txicão e Txucarramãe. Várias espécies de cogumelos silvestres consumidos por povos indígenas da Amazônia já foram registradas, incluindo 34 espécies de cogumelos comestíveis consumidos por diferentes grupos étnicos da Amazônia (Vargas-Isla; Ishikawa; Py-Daniel, 2013). Mais recentemente, (Sanuma *et al.*, 2016) registraram o consumo de 15 espécies de cogumelos pelos Yanomami do extremo norte de Roraima, incluindo espécies dos gêneros *Favolus*, *Hydnopolyphorus*, *Lentinula*, *Lentinus*, *Panus*, *Pleurotus* e *Polyporus*. Em áreas de Mata Atlântica, no entanto, há poucos trabalhos sobre a comestibilidade de macrofungos silvestres, incluindo o consumo de cogumelos dos gêneros *Agaricus*, *Auricularia* e *Macrolepiota* por descendentes de Japoneses, Ucranianos e Italianos que vivem no Sul do Brasil, ou os relatos etnomicológicos sobre o consumo de cogumelos do gênero *Pleurotus* pela etnia Kaigang (Meijer, 2008). Apesar dos poucos registros de cogumelos comestíveis por povos tradicionais em áreas de Mata Atlântica, é nesse bioma que se tem o maior registro de espécies conhecidas de cogumelos para o Brasil, com 2.695 espécies (The Brazilian Flora Group, 2020).

Recentemente, nutricionistas e a população em geral estão se interessando profundamente pelas moléculas bioativas dos cogumelos, considerando o potencial dos cogumelos

Figura 1 – Cogumelos comestíveis

(a) *Pleurotus ostreatus* (Cogumelo Ostra)(b) *Agaricus bisporus* (Champignon de Paris)(c) *Lentinula edodes* (Shiitake)

Fonte: Inaturalist, 2024.

comestíveis como um tesouro inexplorado, que proporciona efeitos benéficos aos humanos em termos de promoção da saúde e redução de riscos relacionados a doenças (Thu *et al.*, 2020b; Bell *et al.*, 2022).

Embora raramente incluídos nas diretrizes alimentares conhecidas, os cogumelos complementam a dieta humana com diversas moléculas bioativas não identificadas ou deficientes em alimentos de origem vegetal e animal, sendo considerados um alimento funcional para a prevenção de diversas doenças humanas (Bell *et al.*, 2022). A busca por recursos naturais alternativos como fontes de espécies químicas (inorgânicas e orgânicas) tem demonstrado interesse, especialmente pela indústria alimentícia.

Recentemente, os cogumelos e outros fungos comestíveis ganharam importância como alimentos funcionais, criando assim uma ponte entre a medicina e a alimentação. Cogumelos contêm uma grande variedade de compostos bioativos, incluindo polissacarídeos, beta-glucanos, triterpenos, esteróis, polipeptídeos e outros antioxidantes. Esses compostos são responsáveis pelos benefícios terapêuticos e medicinais associados aos cogumelos (Bills; Gloer, 2016) (Nandi; Sikder; Acharya, 2019)(Nandi; Sikder; Acharya, 2019)(Chugh *et al.*, 2022). Alguns exemplos incluem:

- Polissacarídeos e beta-glucanos: São carboidratos complexos que desempenham um papel importante no fortalecimento do sistema imunológico. Eles podem estimular a produção de células imunológicas e aumentar a atividade de macrófagos, células natural killer (NK) e linfócitos T. Alguns exemplos de polissacarídeos e beta-glucanos encontrados em cogumelos medicinais incluem lentinana, eritadenina, coriolan e outros.

- Triterpenos e esteróis: São compostos que têm propriedades anti-inflamatórias, antioxidantes e anticancerígenas. Eles podem ajudar a reduzir o estresse oxidativo e a inflamação no corpo, bem como prevenir o crescimento de células cancerígenas. Alguns exemplos de triterpenos e esteróis em cogumelos medicinais incluem ergosterol, ergotioneína, ácido ganodérico, ácidos triterpênicos e outros.

- Polipeptídeos: São proteínas que têm a capacidade de inibir o crescimento de células cancerígenas e vírus. Alguns exemplos de polipeptídeos encontrados em cogumelos medicinais incluem lectinas, lisozimas, ribonucleases e outros.

A crescente demanda por alimentos naturais e saudáveis, bem como a popularidade do segmento gourmet, tem aumentado a aceitação dos cogumelos comestíveis exóticos. Esses alimentos são encontrados em restaurantes finos e mercados especializados, e muitas vezes são vendidos a preços elevados, o que reflete a sua exclusividade e o seu valor agregado.

O mercado mundial de cogumelos comestíveis tem crescido significativamente nos últimos anos. Segundo pesquisa realizada os cogumelos comestíveis, o mercado global de cogumelos comestíveis teve um valor de US\$ 45,3 milhões em 2020 e deve alcançar um valor de US\$ 72,5 bilhões em 2027, impulsionado pelo aumento da demanda por alimentos saudáveis e naturais (Niego *et al.*, 2021). Além disso, destaca o crescente interesse por cogumelos exóticos e de alto valor agregado. Os cogumelos comestíveis são uma excelente fonte de nutrientes e compostos funcionais benéficos para a saúde.

Alguns dos principais benefícios nutricionais e funcionais dos cogumelos comestíveis são:

- Baixo teor calórico: Os cogumelos são baixos em calorias, o que os torna uma excelente escolha para pessoas que desejam controlar a ingestão de calorias. Eles também são ricos em fibras, o que ajuda a manter a sensação de saciedade por mais tempo.

- Fonte de vitaminas e minerais: Os cogumelos são ricos em vitaminas do complexo B, como riboflavina, niacina e ácido pantotênico. Eles também são uma boa fonte de minerais como selênio, cobre e potássio.

- Antioxidantes: Os cogumelos contêm compostos antioxidantes, como ergotioneína e glutationa, que ajudam a combater o estresse oxidativo e a proteger as células dos danos causados pelos radicais livres.

- Regulação do sistema imunológico: Alguns cogumelos, como o shiitake e o mai-take, contêm compostos chamados β -glucanos, que podem ajudar a modular a função imunológica.

- Redução do risco de doenças crônicas: Alguns estudos sugerem que os cogumelos podem ajudar a reduzir o risco de doenças crônicas, como doenças cardíacas, diabetes e câncer, devido aos seus compostos antioxidantes e antiinflamatórios.

- Fonte de proteína: Alguns cogumelos, como os cogumelos portobello e shiitake, são uma boa fonte de proteína e possuem valor nutricional maior do que a maioria das proteínas vegetais (Kalač, 2013).

- Benefícios para a saúde do cérebro: Estudos sugerem que os cogumelos podem ajudar a melhorar a função cognitiva e reduzir o risco de doenças neurodegenerativas, como a doença de Alzheimer (Lee *et al.*, 2019).

- Efeito prebiótico: Os cogumelos contêm fibras prebióticas, que ajudam a promover o crescimento de bactérias benéficas no intestino, melhorando a saúde digestiva.

A produção de cogumelos comestíveis é um processo relativamente simples e de baixo custo, em comparação com outras culturas agrícolas. Por isso, muitos produtores rurais em todo o mundo estão começando a cultivar esses cogumelos como uma fonte alternativa de renda. A produção pode ser feita em pequena escala, utilizando substratos naturais, como resíduos agrícolas, tornando-a uma atividade sustentável e acessível a pequenos produtores. Em resumo, os cogumelos comestíveis apresentam um grande potencial de mercado, especialmente em um contexto de crescente interesse por alimentos naturais, saudáveis e sustentáveis. A crescente demanda por alimentos gourmet e de alto valor agregado, aliada à simplicidade e baixo custo da produção desses cogumelos, torna a sua produção uma oportunidade promissora para produtores rurais e empresas do setor alimentício.

2.1.3 Micronutrientes e outros elementos essenciais para nutrição humana

O ser humano, para a manutenção de suas funções vitais e sobrevivência, requer a ingestão de diversos nutrientes. À medida que seu corpo passa pelos processos de

crescimento, compreendidos como o ciclo de vida, essa necessidade se intensifica. Ao abordarmos essa temática, é possível classificar os nutrientes em dois grupos fundamentais: macronutrientes e micronutrientes.

Os macronutrientes, constituintes essenciais da dieta humana, desempenham papéis cruciais no suporte às funções fisiológicas básicas (Kumar *et al.*, 2017). Estes elementos, em sua maioria, incluem proteínas, carboidratos e lipídios. As proteínas, por exemplo, são fundamentais na construção e reparo dos tecidos do corpo, enquanto os carboidratos fornecem a principal fonte de energia para as atividades diárias (Tappia; Shah, 2022). Por sua vez, os lipídios desempenham um papel vital na absorção de vitaminas lipossolúveis e na manutenção da integridade das membranas celulares.

Em contrapartida, os micronutrientes são necessários em quantidades menores, mas são igualmente essenciais para a saúde e o bem-estar do indivíduo (Arif *et al.*, 2024). Estes compreendem vitaminas e minerais, que desempenham funções específicas em processos metabólicos, enzimáticos e imunológicos. As vitaminas, por exemplo, desempenham papéis-chave como coenzimas em diversas reações bioquímicas, enquanto os minerais contribuem para a saúde dos ossos, função nervosa e equilíbrio hídrico. Por exemplo, segundo Godswill *et al.* (2020), o mineral cálcio representa cerca de 1,5 a 2% do peso corporal total de um humano adulto.

O zinco, por exemplo, é crucial como cofator em reações enzimáticas e na regeneração de órgãos. O magnésio controla o sistema neuroendócrino e participa do metabolismo do cálcio, enquanto o potássio é essencial para o equilíbrio eletrolítico. No entanto, o consumo excessivo de micronutrientes pode levar a consequências prejudiciais, como intoxicação por vitaminas solúveis em gordura (A, D, E e K) e distúrbios gastrointestinais por minerais como zinco, selênio e ferro. Assim, é crucial manter um equilíbrio adequado na ingestão desses nutrientes para garantir uma boa saúde (Arif *et al.*, 2024).

Assim, a ingestão de macronutrientes e micronutrientes se torna importante para garantir a sobrevivência e, além disso, a qualidade de vida do ser humano ao longo da vida. A nutrição adequada, baseada na compreensão desses elementos essenciais, desempenha um papel fundamental na promoção da saúde e na prevenção de doenças associadas à alimentação desequilibrada. Alguns micronutrientes e elementos essenciais à saúde humana e suas funções podem ser vistos no Quadro 1.

Quadro 1 – Micronutrientes e elementos essenciais

| Elementos | Função |
|---|--|
| Micronutrientes | |
| Cálcio (ca) | Ajuda na formação óssea, principalmente em crianças e na regulação metabólica. |
| Cromo (Cr) | Potencializa a ação da insulina nos tecidos periféricos e intervém no metabolismo dos carboidratos, proteínas, gorduras e no estado oxidativo. |
| Cobalto (Co) | Ajuda na formação da vitamina B12 (hidroxocobalamina) e é essencial para a produção e maturação dos glóbulos vermelhos no sangue. |
| Cobre (Cu) | Ajuda na produção de energia, ajuda no metabolismo do ferro, formação do pigmento melanina, hormônio da tireoide e o metabolismo da glicose |
| Flúor (F) | Ajuda na saúde bucal e óssea. |
| Iodo (I) | Essencial para os hormônios da tireoide e regulador da função da glândula tireoide. |
| Ferro (Fe) | Função de transporte de oxigênio, ativação do oxigênio molecular e transporte de elétrons. |
| Manganês (Mn) | Regulação do açúcar no sangue e da energia celular, reprodução, digestão, crescimento ósseo, coagulação sanguínea e hemostasia, defesa antioxidant e função imunológica. |
| Molibdênio (Mo) | Redução de nitratos, destoxificação de sulfitos, catabolismo de purinas e redução de substratos N-hidroxilados. |
| Selênio (Se) | Atividade antioxidant e redox, controle do metabolismo dos hormônios tireoidianos. |
| Zinco (Zn) | Fundamental para o controle da transcrição do DNA em RNA e sistema de defesa antioxidant. |
| Elementos essenciais para a saúde humana | |
| Fósforo (P) | Ajuda na produção de ATP (adenosina trifosfato), mantém os ossos e dentes fortes e saudáveis. |
| Níquel (Ni) | Estabiliza as membranas celulares e a estrutura do DNA e RNA. |
| Sódio (Na) | Ajuda na transmissão de impulsos nervosos |

Fonte: Retirada de (Berger *et al.*, 2022; Cozzolino, 2015; Zoltán, 2019)

2.1.3.1 Cogumelos como fonte de micronutrientes e outros elementos essenciais para nutrição humana

Os cogumelos possuem transportadores de íons nas membranas celulares que facilitam a captação de metais do ambiente. Além disso, os cogumelos podem secretar moléculas orgânicas, como ácidos orgânicos, que formam complexos estáveis com os íons metálicos, auxiliando na absorção desses elementos essenciais. A biossíntese de ligantes específicos, como sideróforos, também é uma estratégia utilizada pelos cogumelos para se ligarem a íons metálicos de alta afinidade. Alguns desses íons metálicos atuam como cofatores para enzimas nos cogumelos, sendo essenciais para processos metabólicos, como a produção de ácidos orgânicos. Essa interação entre cogumelos e metais é fundamental para a compreensão dos processos metabólicos desses organismos e tem implicações importantes na biotecnologia (Karaffa; Fekete; Kubicek, 2021).

Além disso, a regulação dos transportadores de íons e a expressão de genes relacionados à captação e metabolismo de metais são influenciadas pela disponibilidade de íons metálicos no ambiente. Os cogumelos podem ajustar sua resposta de acordo com a concentração e o tipo de metal presente, garantindo a homeostase metálica intracelular. Essa capacidade de adaptação dos cogumelos a diferentes condições de metal contribui para sua sobrevivência e desempenho metabólico em ambientes variados (Priyadarshini *et al.*, 2021).

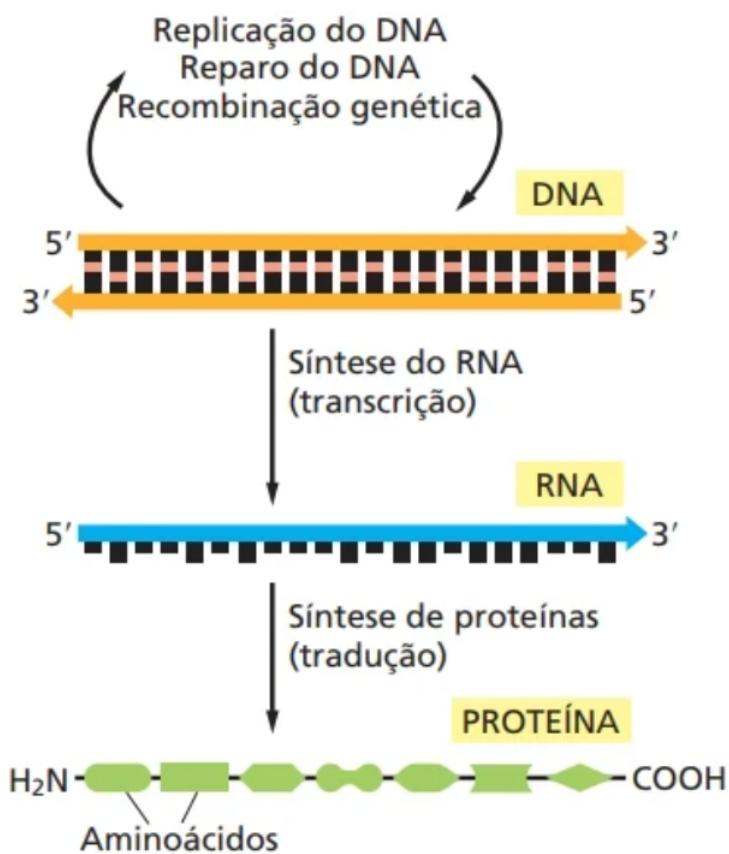
Na pesquisa sobre o papel dos íons de metais na acumulação de ácidos orgânicos por cogumelos, foram testados vários metais, incluindo manganês (Mn), ferro (Fe), cobre (Cu) e zinco (Zn). A interação entre cogumelos e íons metálicos desempenha um papel crucial em processos biotecnológicos, como a produção de ácidos orgânicos por fermentação. A pesquisa conduzida por Karaffa, Fekete e Kubicek (2021) investigou o impacto dos íons metálicos na acumulação de ácidos orgânicos por cogumelos, abrangendo uma gama de metais. Dentre os metais avaliados, incluíram-se ferro (Fe), cobre (Cu) e zinco (Zn), que são micronutrientes essenciais para a saúde humana, e manganês (Mn), que também é um elemento essencial. Compreender os mecanismos pelos quais os cogumelos se ligam a íons de metais é essencial para otimizar o estudo dessa interação, fornecendo *insights* valiosos para o desenvolvimento de estratégias de produção de cogumelos ricos em nutrientes essenciais para nutrição humana.

2.1.4 Genômica de Fungos

O dogma central da biologia molecular é um conceito essencial que descreve o fluxo unidirecional da informação genética dentro das células. Ele estabelece que o DNA é replicado para preservar a informação genética durante a divisão celular, que esta informação é transcrita do DNA para RNA, e que o mRNA é então traduzido em proteínas (Alberts *et al.*, 2017). Este paradigma demonstrado na Figura 2, esclarece como os

organismos mantêm e transmitem informações genéticas, adaptando sua expressão genética em resposta à condições ambientais variáveis e desafios biológicos específicos. Todas as células, desde bactérias, fungos até seres humanos, expressam sua informação genética dessa maneira (Alberts *et al.*, 2017). Para os fungos, assim como para todos os outros organismos, esse dogma é essencial para a expressão e regulação dos genes. A replicação do DNA assegura a hereditariedade estável das características, garantindo que a informação genética seja transmitida corretamente às células filhas durante a divisão celular (Jain; Singh, 2009).

Figura 2 – Dogma Central da Biologia Molecular



Fonte: Retirado de (Alberts *et al.*, 2017)

Na genômica fúngica, o estudo do genoma completo dos fungos revela *insights* cruciais sobre sua organização genética, diversidade e adaptação. Através da transcrição, os fungos ativam ou desativam genes em resposta a estímulos ambientais, ajustando sua fisiologia e bioquímica para enfrentar desafios específicos.

Basidiomycota é um filo de fungos e cogumelos com grande importância econômica e ecológica, desempenhando um papel crucial nos ecossistemas (Bisanção; Polonio; Golias, 2022). Esses organismos possuem genomas com uma notável variedade de tamanho, variando de algumas dezenas de milhões a vários bilhões de pares de bases (Mohanta; Bae,

2015). O tamanho médio do genoma das espécies de cogumelos Basidiomycota é de 46,48 Mb (megapares de bases), o que corresponde a 46.480 milhões de pares de bases de DNA (Mohanta; Bae, 2015).

A complexidade genômica dos cogumelos do filo *Basidiomycota* representa um desafio na busca por genes de interesse. A presença de material genético não codificante e repetições de sequências dificulta a identificação e isolamento de genes específicos (Yuan *et al.*, 2019). No entanto, avanços nas tecnologias de NGS e nas abordagens bioinformáticas têm impulsionado a pesquisa nessa área (Larson *et al.*, 2022). A identificação e caracterização de genes de importância econômica, ecológica e medicinal nos cogumelos do filo *Basidiomycota* têm implicações significativas em diversas áreas, como nutrição, biotecnologia, ecologia, medicina e agricultura.

A análise dos genomas fúngicos é uma ferramenta fundamental para compreender a diversidade, evolução e adaptação desses organismos. O sequenciamento de genomas de fungos tem proporcionado um melhor entendimento de sua biologia e fisiologia, incluindo a identificação de *clusters* de genes biosintéticos e genes conservados entre espécies (Keller, 2019). Essa análise tem sido relevante na identificação de genes de interesse biotecnológico, como aqueles relacionados à produção de enzimas lignocelulolíticas, que desempenham um papel importante na produção de biocombustíveis e biorremediação (Saini; Sharma, 2021).

Os genes fúngicos desempenham um papel essencial na regulação dos processos metabólicos, crescimento, reprodução e resposta a estímulos ambientais. Eles são responsáveis pela produção de enzimas, proteínas estruturais e outros componentes essenciais para o funcionamento dos fungos. A análise desses genes permite identificar e compreender as vias metabólicas envolvidas na síntese de compostos bioativos, na resposta a estresses ambientais e na interação com outros organismos.

Alguns estudos recentemente realizados identificaram informações importantes sobre cogumelos comestíveis através de seus genomas. A pesquisa em genômica está revelando dados cruciais sobre a composição nutricional e o potencial alimentício de espécies como *Stropharia rugosoannulata* e *Hericium erinaceus*. Estes cogumelos têm sido extensivamente estudados, oferecendo insights valiosos sobre suas características genéticas e aplicações na alimentação. Por exemplo, o sequenciamento do genoma de *Stropharia rugosoannulata* identificou uma variedade impressionante de 12.752 genes codificadores de proteínas, incluindo 486 genes de enzimas CAZyme, fundamentais para processos metabólicos e degradação de lignocelulose (Li *et al.*, 2022b; Gong *et al.*, 2020).

Hericium erinaceus, conhecido como Lion's Mane, também foi objeto de análises genômicas detalhadas, resultando na identificação de 10.620 genes preditos e 341 CAZymes. Esses dados não apenas destacam sua capacidade única de degradar materiais vegetais, mas também elucidam seus potenciais benefícios nutricionais e medicinais, incluindo propriedades neuroprotetoras e antioxidantes associadas a compostos como erinacinas e

hericenonas que são importantes no processo de crescimento neural (Gong *et al.*, 2020).

2.1.4.1 *Proteínas transportadoras*

As proteínas são macromoléculas essenciais para a vida, compostas por uma sequência de aminoácidos conectados por ligações peptídicas. Elas desempenham uma ampla variedade de funções nos organismos vivos, desde estruturar e fortalecer tecidos até catalisar reações químicas e transportar substâncias (Chen; Dong; Minteer, 2020). Existem diversos tipos de proteínas, cada uma com função específica: algumas são estruturais, como a queratina da pele e cabelo, enquanto outras são enzimas que aceleram reações bioquímicas (Liu; Dong, 2020). Além de atuar como hormônios, transportadores, defensores imunológicos e componentes musculares contráteis, sua diversidade funcional é crucial para o organismo.

As proteínas são consideradas as "operárias" celulares, formadas por cadeias de aminoácidos dobradas em estruturas tridimensionais específicas que conferem propriedades e funções distintas. Por exemplo, proteínas envolvidas no transporte de elétrons desempenham papel vital (Cosic; Cosic; Loncarevic, 2020). Interagem com outras moléculas através de ligações químicas como hidrogênio e hidrofóbicas, o que permite funções específicas em diferentes contextos (Chen; Dong; Minteer, 2020). Enzimas catalisam reações, transportadores movem substâncias através de membranas e fatores de transcrição regulam a expressão gênica (Cosic; Cosic; Loncarevic, 2020; Liu; Dong, 2020).

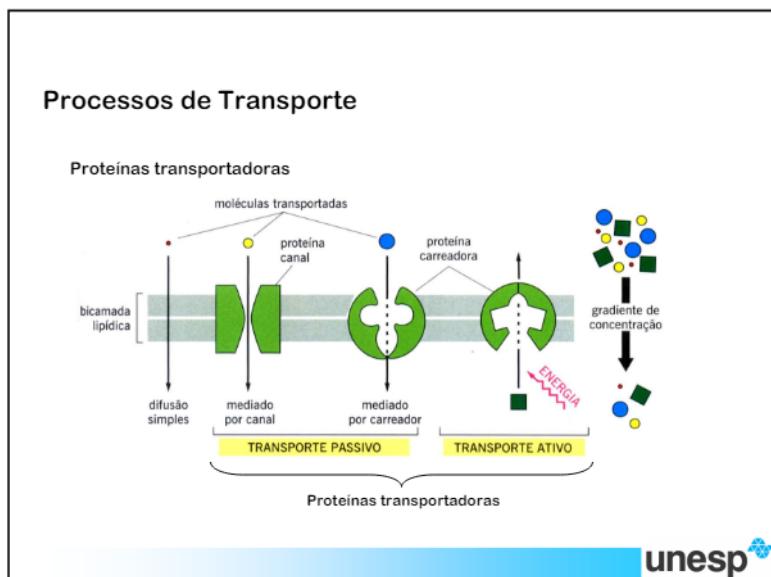
Proteínas são essenciais na estrutura, função e regulação dos seres vivos (Aptekmann *et al.*, 2022), participando desde transporte de nutrientes até defesa imunológica. Modificações pós-traducionais como fosforilação e glicosilação afetam suas funções, regulando processos biológicos em resposta a sinais como estímulos hormonais. Compreender seu papel é crucial para avanços em biologia e medicina, contribuindo para terapias direcionadas (Whitford, 2013).

Já as proteínas transportadoras são responsáveis por manter o equilíbrio da célula. São responsáveis por transportar substâncias específicas através das membranas celulares. Estas substâncias podem incluir íons, nutrientes, resíduos essenciais para a sobrevivência e funcionamento da célula (Permyakov, 2021). A capacidade dessas proteínas de mover substâncias para dentro e para fora da célula é vital para a manutenção da homeostase celular, o equilíbrio necessário para a saúde e a atividade celular. Além disso, as proteínas transportadoras representam aproximadamente 25% das proteínas codificadas pelo genoma, destacando sua importância e abundância nos processos biológicos (Tusnády; Dosztányi; Simon, 2004).

Existem diferentes tipos de proteínas transportadoras, cada uma especializada em transportar tipos específicos de moléculas. As proteínas de transporte ativo, por exemplo, utilizam energia para mover substâncias contra seus gradientes de concentração

(Bartley; Davies, 1954), enquanto as proteínas de transporte passivo permitem a difusão de substâncias ao longo desses gradientes (Parker; Dunham, 2020)(Figura 3). Exemplos de proteínas transportadoras incluem as bombas de íons, que mantêm o equilíbrio eletrolítico, e as proteínas transportadoras de glicose, que regulam os níveis de glicose no sangue. A disfunção dessas proteínas pode levar a várias doenças e condições, como distúrbios neurológicos (Gil-Martins *et al.*, 2020). Além de sua função essencial no transporte de substâncias, as proteínas transportadoras também desempenham papéis importantes na comunicação entre células e na resposta a sinais externos (Zhou *et al.*, 2023).

Figura 3 – Processo de Transporte

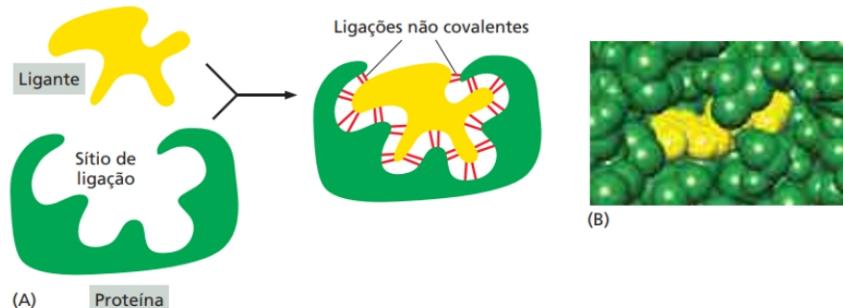


Fonte: Fonte: Prof Cesar Martin (UNESP)

2.1.4.2 Proteínas que se ligam a íons de metais

Entre as muitas capacidades das proteínas, como a ligação a substratos e a outras proteínas, destaca-se a habilidade de algumas em se ligar a metais através dos íons metálicos. Todas as proteínas possuem a capacidade de se ligar a outras moléculas, e qualquer substância que se ligue a proteínas é denominada ligante. É importante ressaltar que essa ligação só ocorre se a molécula do ligante se ajustar à proteína em um processo análogo a um encaixe, conforme ilustrado na Figura 4. Duas classes importantes de proteínas que realizam esse processo são as proteínas de canais iônicos e as proteínas carreadoras (Alberts *et al.*, 2017).

Figura 4 – Processo de ligação de proteínas



Fonte: Retirado de (Alberts *et al.*, 2017)

Os canais iônicos são proteínas que formam poros na membrana celular, permitindo a passagem de íons específicos de um lado para o outro da membrana. Essas proteínas são essenciais para a manutenção do potencial de membrana. A especificidade dos canais iônicos é determinada pela estrutura tridimensional do canal, que permite a passagem apenas de íons com determinadas características, como tamanho e carga elétrica (Alberts *et al.*, 2002).

As proteínas carreadoras, por outro lado, são responsáveis pelo transporte ativo ou passivo de moléculas através da membrana celular. Essas proteínas se ligam a moléculas específicas, como nutrientes ou íons, e mudam de conformação para transportar essas moléculas através da membrana. No caso do transporte ativo, essa mudança de conformação é acoplada à hidrólise de ATP, fornecendo a energia necessária para o transporte contra um gradiente de concentração (Alberts *et al.*, 2002).

Os metais frequentemente encontrados em complexos proteicos incluem zinco, ferro, cobre, manganês e magnésio (Permyakov, 2021). Cada um desses metais desempenha papéis específicos nas proteínas. Por exemplo, o zinco é essencial para a atividade catalítica de muitas enzimas (Alberts *et al.*, 2002; Permyakov, 2021).

2.2 REFERENCIAL COMPUTACIONAL

Nesta seção, serão apresentados os referenciais computacionais envolvidos neste projeto. Dentre os assuntos abordados, destacam-se os conceitos e aplicações da Bioinformática, o pipeline em Bioinformática, a mineração em genomas, as redes neurais, Python e BioPython, além de Celery e RabbitMQ.

2.2.1 Bioinformática

A bioinformática combina a biologia e a ciência da computação, utilizando técnicas e ferramentas computacionais para analisar e interpretar dados biológicos. Sua origem

remonta à década de 1960, quando o sequenciamento de DNA começou a gerar grandes volumes de dados que exigiam processamento eficiente (Martí-Carreras; Maes, 2019). Desde então, a bioinformática tem desempenhado um papel fundamental no avanço da pesquisa biológica, proporcionando abordagens e ferramentas para o estudo de genomas, proteínas, expressão gênica e outras áreas da biologia molecular.

Com o advento do NGS, a geração de dados biológicos se tornou ainda mais massiva e complexa, exigindo abordagens computacionais avançadas para análise e interpretação (Srivastava; Naik, 2021). A bioinformática evoluiu para incorporar métodos de análise de sequências genômicas, modelagem de proteínas, predição de estrutura e função, além da integração de dados de diferentes fontes.

As aplicações da bioinformática são amplas e impactam várias áreas da pesquisa biológica e biomédica. No campo da genômica, a bioinformática desempenha um papel fundamental na montagem e anotação de genomas, identificação de genes e elementos regulatórios, análise de variação genética e comparação entre diferentes genomas. Na área de proteômica, a bioinformática contribui para a identificação de proteínas, predição de sua estrutura e função, análise de interações proteína-proteína e identificação de alvos terapêuticos. Por exemplo Soleymani *et al.* (2023) desenvolveu uma arquitetura que utiliza Aprendizado Profundo, para fazer predição de interações proteína-proteína. Além disso, a bioinformática é crucial na análise de dados de expressão gênica, permitindo a identificação de genes diferencialmente expressos, vias metabólicas e construção de redes de interação gênica. Também é utilizada na epidemiologia molecular, análise de metagenômica e descoberta de biomarcadores em doenças complexas.

2.2.1.1 Pipeline de Bioinformática

A utilização de pipelines na construção de sistemas de bioinformática desempenha um papel essencial na análise de dados biológicos. Uma pipeline consiste em uma sequência de etapas interconectadas, cada uma executando uma função específica, por exemplo, pré-processamento, alinhamento de sequências, anotação genômica ou análise estatística (Wratten; Wilm; Göke, 2021).

A implementação de pipelines traz diversas vantagens para a bioinformática. Em primeiro lugar, elas permitem a padronização e a reprodução de análises, garantindo que os mesmos procedimentos sejam aplicados consistentemente a diferentes conjuntos de dados. Além disso, as pipelines facilitam a gestão de grandes volumes de dados, pois as etapas são automatizadas e executadas de forma sequencial, reduzindo erros e otimizando o processamento (Bourgey *et al.*, 2019). A integração de diferentes ferramentas e algoritmos em uma pipeline viabiliza a combinação de métodos complementares, possibilitando a obtenção de resultados mais abrangentes e confiáveis.

2.2.1.2 Mineração em genomas

A mineração de dados é uma disciplina que busca descobrir padrões, tendências e informações relevantes em conjuntos de dados complexos, utilizando técnicas estatísticas, matemáticas e computacionais. Nos últimos anos, a mineração de dados genômicos tem ganhado destaque devido aos avanços na tecnologia de sequenciamento genético e à geração massiva de informações genéticas armazenadas em bancos de dados. O estudo proposto por (Yeh; Yeh; Chen, 2022) utilizou a mineração de dados genômicos para criar uma rede genética e epigenética de todo o genoma humano (GWGEN), avanços como esse pode melhorar ainda mais o conhecimento sobre sobre a vida humana.

Além disso, outros estudos demonstram a importância da utilização de técnicas de mineração em genomas. Por exemplo, o estudo proposto por (Bauman *et al.*, 2021) dedicou-se à mineração de dados genômicos para identificar compostos bioativos, dividindo-se em mineração de informações dos compostos e mineração de informações dos grupos. No entanto, a mineração de dados genômicos enfrenta desafios únicos devido à imensa quantidade de dados gerados, exigindo uma infraestrutura computacional robusta, algoritmos eficientes de processamento e técnicas que facilitem a obtenção dessas informações. A revisão realizada por (Su *et al.*, 2020) apresenta dados sobre métodos computacionais utilizados para minerar genomas e suas limitações.

A análise de informações genéticas por meio da mineração de dados pode fornecer *insights* valiosos sobre a estrutura do DNA, a expressão gênica, a presença de variantes genéticas e sua associação com doenças, entre outros aspectos relacionados à genética (Florez, 2017). Essa área de estudo possibilita a extração de informações significativas a partir de dados complexos e confusos, contribuindo para avanços no campo da genética e medicina.

Uma das principais aplicações da análise de dados genômicos é a identificação de variantes genéticas associadas a doenças, como Parkinson (Consortium *et al.*, 2011). Com o avanço da tecnologia de sequenciamento genético, é possível analisar o genoma de um indivíduo em busca de alterações genéticas que possam predispor a certas condições de saúde (Liu *et al.*, 2022). Essa abordagem, conhecida como estudo de associação genômica, permite a descoberta de marcadores genéticos relevantes para diagnóstico, prognóstico e desenvolvimento de terapias personalizadas.

Além disso, a análise de dados genômicos também é fundamental para a compreensão dos mecanismos moleculares subjacentes a processos biológicos complexos, como a identificação de assinaturas biosintéticas em dados genômicos, o que leva a um entendimento sobre a biologia de produtos naturais, química e descoberta de medicamentos (Prihoda *et al.*, 2021).

Ao analisar o genoma de um indivíduo, é possível obter informações sobre suas

predisposições genéticas a certas condições de saúde, sua resposta a determinados medicamentos e até mesmo suas características individuais, como metabolismo de medicamentos. Essas informações podem ser usadas para orientar decisões clínicas (Kwan *et al.*, 2020).

Mas a pergunta mais importante é: “Como fazer a mineração de dados em genoma ?” A mineração em genomas abrange diversos métodos, como a análise de variantes genéticas, as GWAS, a análise de expressão gênica, as redes regulatórias e a análise de enriquecimento de vias (Bristot, 2022). Além disso, o uso de redes neurais e Aprendizado Profundo tem se destacado nesse campo (Zhang *et al.*, 2019). Essas técnicas são capazes de reconhecer padrões complexos em grandes volumes de dados genômicos, permitindo a classificação de variantes genéticas, a previsão de estruturas de proteínas e a inferência filogenética, entre outros. A aplicação de redes neurais convolucionais e recorrentes tem impulsionado a descoberta de associações genéticas, a previsão de fenótipos e a classificação de pacientes com base em informações genômicas (Wlodarczak; Soar; Ally, 2015; Chu *et al.*, 2020). Essas abordagens promissoras abrem novas perspectivas para a compreensão do genoma de qualquer organismo e possibilita o desenvolvimento novas soluções.

Para obter essas informações, dentre muitos outros métodos, podemos utilizar as redes neurais e técnicas de aprendizado de máquina (Prihoda *et al.*, 2021). Esses métodos avançados permitem analisar grandes volumes de dados biológicos e clínicos para identificar padrões. As redes neurais, por exemplo, podem ser treinadas para prever deficiências nutricionais ou doenças relacionadas com base em perfis dietéticos e genéticos (Dutta *et al.*, 2020).

2.2.2 Redes Neurais Artificiais

As redes neurais artificiais, inspiradas no cérebro humano, são modelos computacionais compostos por unidades interconectadas, conhecidas como neurônios artificiais ou "nodes". Esses modelos têm sido amplamente utilizados em diversas áreas, como reconhecimento de padrões, processamento de linguagem natural, visão computacional e aprendizado de máquina (Liao *et al.*, 2021).

Os neurônios artificiais são a base fundamental de uma rede neural. Cada neurônio recebe uma entrada, realiza um cálculo ponderado e produz uma saída. Esses neurônios são organizados em camadas, incluindo a camada de entrada, camadas ocultas e camada de saída. As conexões entre os neurônios são ponderadas por valores chamados de pesos sinápticos, que determinam a influência que um neurônio exerce sobre o outro (Yang; Wang, 2020).

A rede neural funciona em duas etapas principais: propagação direta e retropropagação do erro. Durante a propagação direta, os dados de entrada são passados pela rede, camada por camada, até alcançar a camada de saída. Cada neurônio processa as

informações recebidas e as transmite para a próxima camada. A pesquisa desenvolvida por Yang e Wang (2020), explicou esse conceito de forma mais detalhada, demonstrando o seu funcionamento em relação a um neurônio humano. O cérebro humano recebe uma informação, a interpreta e resulta em uma ação. No neurônio artificial, a função de ativação de cada neurônio introduz não-linearidades no processamento dos dados, permitindo a modelagem de relações complexas.

A função de ativação desempenha um papel crucial nas redes neurais artificiais, sendo um componente fundamental. Sua principal função é introduzir não-linearidade nas saídas dos neurônios, viabilizando que a rede neural tenha a capacidade de aprender e representar relações complexas nos dados de entrada (Gomes; Ludermir, 2021).

O aprendizado em uma rede neural é frequentemente realizado através do algoritmo de repropagação. Durante o treinamento, a rede compara a saída produzida com o resultado esperado, calculando o erro associado. Esse erro é então propagado de volta pela rede, ajustando gradualmente os pesos sinápticos para minimizar o erro. Esse processo de ajuste dos pesos permite que a rede aprenda a realizar tarefas específicas e reconhecer padrões (Kopiler *et al.*, 2019, 2019; Matos¹ *et al.*, 2021).

Existem várias arquiteturas de redes neurais, cada uma com suas características específicas. As redes neurais *feedforward* são as mais comuns, onde a informação flui em uma única direção, da camada de entrada para a camada de saída. Redes neurais convolucionais (CNNs) são amplamente utilizadas para processamento de imagens (Coudray *et al.*, 2018), enquanto redes neurais recorrentes (RNNs) são mais adequadas para análise de sequências temporais (Hewamalage; Bergmeir; Bandara, 2021). As redes neurais profundas, também conhecidas como *Deep Neural Networks*, possuem múltiplas camadas ocultas e são capazes de aprender representações hierárquicas de dados complexos (Souza *et al.*, 2020). As redes neurais possuem a configuração de 3 (três) conjuntos interligados:

- Camada de entrada

A camada de entrada é responsável por receber os dados de entrada e transmiti-los para a próxima camada da rede neural. Cada nó (ou neurônio) na camada de entrada representa uma característica (ou atributo) do dado de entrada.

- Camada oculta

Camadas ocultas são camadas intermediárias entre a camada de entrada e a camada de saída. Elas são chamadas de “ocultas” porque seus neurônios não estão diretamente conectados às entradas ou saídas do sistema. Uma rede neural pode ter uma ou várias camadas ocultas. Cada neurônio em uma camada oculta recebe os sinais de entrada de neurônios na camada anterior, executa uma transformação não linear nesses sinais e transmite os resultados para a próxima camada.

- Camada de saída

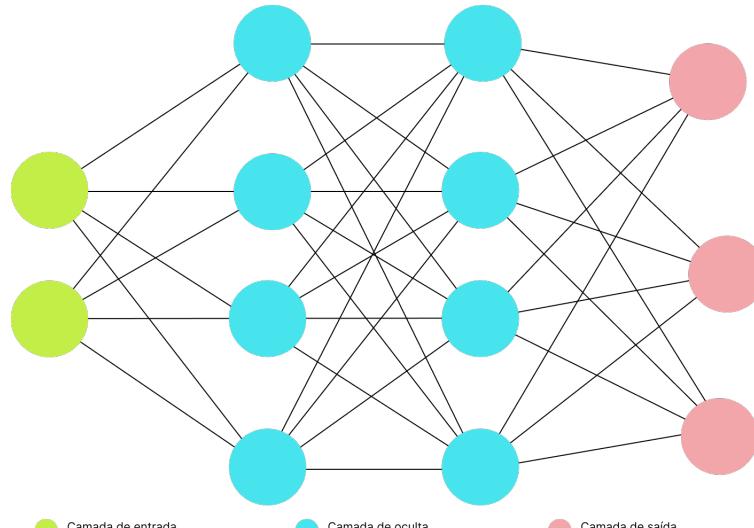
A camada de saída é responsável por produzir a saída final da rede neural. Ela recebe os sinais dos neurônios na última camada oculta e produz a previsão ou resultado desejado. O número de neurônios na camada de saída depende do problema em questão.

2.2.2.1 Modelos de Redes Neurais

Dentre os inúmeros modelos disponíveis, abordaremos especificamente três deles.

- Redes Neurais Convolucionais (CNN), demonstrado na Figura 5, são arquiteturas de redes neurais especializadas em processar dados de grade, como imagens. Elas aplicam operações de convolução e *pooling* para extrair características relevantes e reduzir a dimensionalidade dos dados. As CNNs são amplamente utilizadas em visão computacional e reconhecimento de padrões, sendo capazes de identificar informações complexas em imagens e impulsionar avanços em diversas áreas, como medicina e indústria automotiva (Olivatto, 2021).

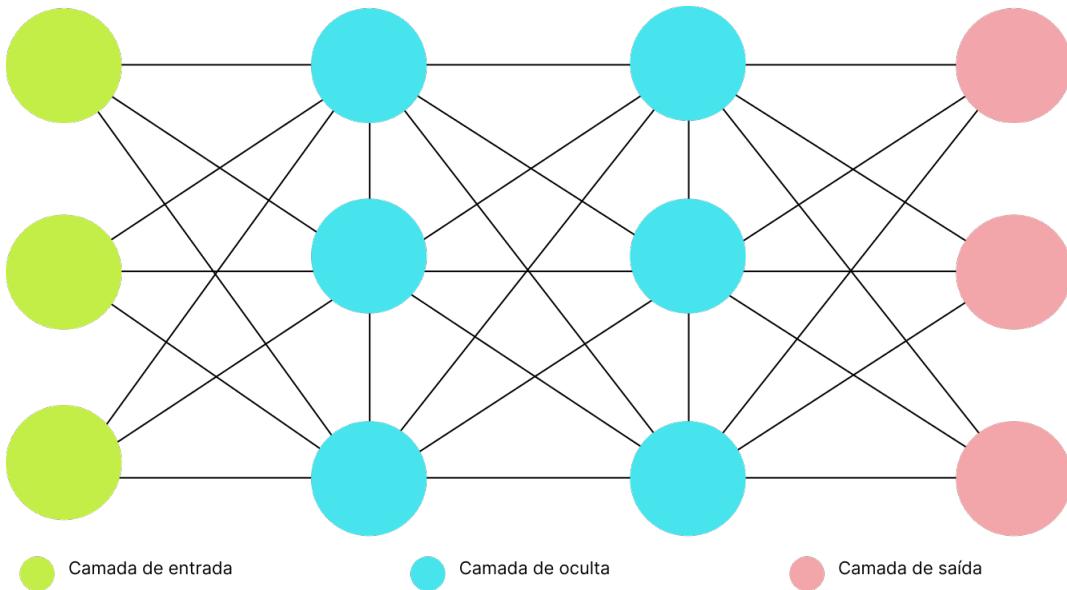
Figura 5 – Redes neural convolucional.



Fonte: Produzido pelo autor (2024)

- Redes Neurais Recorrentes (RNN), Figura 6, são arquiteturas de redes neurais que processam dados sequenciais utilizando conexões de retroalimentação. Elas possuem células de memória que permitem capturar dependências temporais e contextuais em sequências de dados. As RNNs são amplamente aplicadas em tarefas como processamento de linguagem natural, análise de séries temporais e geração de texto, devido à sua capacidade de modelar o histórico completo da sequência e capturar padrões e contextos relevantes (Wang *et al.*, 2019).

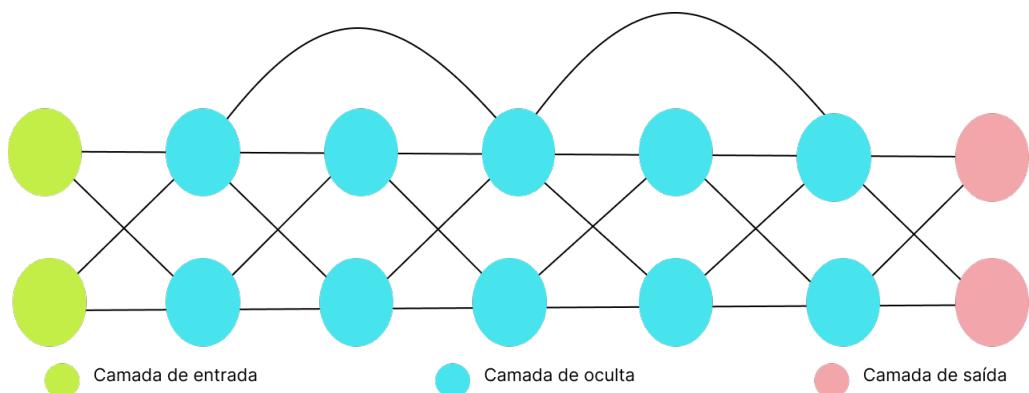
Figura 6 – Redes Neurais Recorrentes



Fonte: Produzido pelo autor (2024)

- A Deep Residual Network (DRN) apresentada na Figura 7 é uma arquitetura de rede neural que utiliza blocos residuais para treinar redes profundas de forma eficiente. Esses blocos permitem que a rede aprenda as diferenças residuais entre as ativações esperadas e as reais, evitando a degradação do desempenho à medida que a rede fica mais profunda. O DRN tem sido bem-sucedido em tarefas como classificação de imagens, detecção de objetos e segmentação semântica, estabelecendo-se como uma arquitetura popular no campo do aprendizado profundo (Tang; Wu, 2019).

Figura 7 – Deep Residual Network.



Fonte: Produzido pelo autor (2024)

2.2.2.2 Aprendizado Profundo (Deep Learning)

Aprendizado Profundo é uma técnica que se baseia em redes neurais artificiais profundas, compostas por camadas de unidades de processamento interconectadas, inspiradas na estrutura do cérebro humano. Essas redes são capazes de aprender e extrair caracterís-

ticas complexas dos dados, permitindo que os sistemas desenvolvam uma compreensão mais sofisticada e abstrata das informações fornecidas.

Uma das principais vantagens do Aprendizado Profundo é sua capacidade de lidar com grandes volumes de dados e encontrar padrões sutis que podem passar despercebidos por outros métodos. Isso tem impulsionado avanços significativos em áreas como reconhecimento de fala, visão computacional (Olivatto, 2021), tradução automática e processamento de linguagem natural (Khan; Abid; Abid, 2020).

O desenvolvimento de assistentes virtuais inteligentes, como a Siri e o Google Assistant, por exemplo, foi possibilitado pelo uso de técnicas de Aprendizado Profundo (Someshwar *et al.*, 2020). O aprendizado de máquina profundo é amplamente utilizado na Bioinformática devido à sua capacidade de processar grandes quantidades de dados biológicos para classificar e extrair informações relevantes para a pesquisa nessa área (Cao *et al.*, 2020). A principal diferença entre redes neurais e aprendizado profundo é a complexidade e a profundidade das camadas ocultas. Redes neurais são modelos complexos que simulam artificialmente o funcionamento do cérebro humano (Yang; Wang, 2020), podendo ter uma ou várias camadas ocultas. O aprendizado profundo, por outro lado, utiliza redes neurais com muitas camadas ocultas para realizar inferências a partir dos dados.

2.2.3 Aplicações Web para Bioinformática

As aplicações web em bioinformática têm se destacado como ferramentas acessíveis e eficientes para a análise e interpretação de dados biológicos (Oliveira; Sundfeld, 2022; Mendes *et al.*, 2022; Waldemarin, 2021). As aplicações web oferecem uma conveniência significativa ao fornecer acesso imediato e colaborativo a uma ampla gama de recursos e ferramentas bioinformáticas, eliminando a necessidade de instalação local e conhecimentos avançados em programação. Esse tipo de aplicação permite que pesquisadores de diversas áreas possam utilizar ferramentas complexas e realizar análises sofisticadas sem a barreira de configurar e manter softwares específicos em seus computadores, necessitando apenas de um navegador (browser) e conexão com a internet.

Além de simplificar o acesso, as aplicações web promovem a colaboração entre pesquisadores ao facilitar o compartilhamento de dados e resultados em tempo real. Isso possibilita que equipes localizadas em diferentes partes do mundo trabalhem juntas de maneira mais eficiente e integrada, acelerando o progresso das pesquisas e a disseminação de conhecimentos. A natureza centralizada dessas plataformas também garante que todos os usuários tenham acesso às versões mais recentes das ferramentas, reduzindo incompatibilidades e garantindo que as análises sejam realizadas com os métodos mais atualizados disponíveis (Oliveira; Sundfeld, 2022). Um exemplo notável de aplicação web em bioinformática é o PA-Star-Web, que permite o alinhamento múltiplo ótimo de sequências

biológicas através de um servidor web acessível (Mendes *et al.*, 2022). Essa abordagem simplifica a análise de sequências, contribuindo para o entendimento de relações evolutivas e investigação de padrões genéticos. Outra aplicação relevante é a OLATCG, que utiliza a web como plataforma educacional para o ensino de genética no ensino médio (Oliveira; Sundfeld, 2022). Essa ferramenta oferece procedimentos experimentais *in silico* e interação com ferramentas de bioinformática, aproximando os alunos da pesquisa científica.

Além disso, o desenvolvimento baseado em modelos de serviços adaptadores tem sido uma abordagem promissora para a integração de ferramentas bioinformáticas em servidores web (Waldemarin, 2021). Esses modelos adaptadores garantem a compatibilidade e a padronização da interação entre as ferramentas e o usuário, simplificando o acesso e ampliando a gama de recursos disponíveis para análise bioinformática. Considerando que essas aplicações não demandam grandes recursos computacionais, o tempo de execução pode ser otimizado significativamente (Elbaum; Karre; Rothermel, 2003). Isso é possível porque a maioria das tarefas bioinformáticas envolvidas são computacionalmente leves, permitindo uma execução rápida e eficiente, mesmo em servidores web com recursos limitados. Como resultado, os usuários podem realizar análises complexas em menor tempo, aumentando a produtividade e a eficiência no processamento de dados bioinformáticos (Messina *et al.*, 2018).

2.2.3.1 Python e Biopython

Python desempenha um papel fundamental na área da bioinformática devido à sua versatilidade e facilidade de uso. Com uma vasta gama de bibliotecas e ferramentas especializadas, Python se tornou a escolha preferida para a análise de dados biológicos, sendo utilizada na grande maioria dos aplicativos encontrados já existentes (Uzma *et al.*, 2022; Serrano, 2020; Soleimani *et al.*, 2023; Yang; Yang; Tu, 2022; Monaco *et al.*, 2021; Hannigan *et al.*, 2019). Através de bibliotecas renomadas, como Biopython, pandas e NumPy, os pesquisadores podem manipular, analisar e visualizar conjuntos de dados genômicos de grandes proporções, explorando informações valiosas.

Com bibliotecas específicas para a manipulação de sequências, alinhamentos e estruturas de proteínas, Python é uma linguagem de programação poderosa, com bibliotecas que podem ser utilizadas para a interpretação de dados biológicos, como a biblioteca Biopython (Cock *et al.*, 2009).

Biopython é uma biblioteca de código aberto amplamente utilizada na área de bioinformática e biologia computacional. Desenvolvida em Python, ela fornece um conjunto abrangente de ferramentas e módulos para a manipulação, análise e visualização de dados biológicos (Cornish; Kricka; Park, 2021).

Com Biopython, é possível realizar tarefas bioinformáticas como manipulação de sequências de DNA, RNA e proteínas, alinhamentos de sequências, análise de estruturas

proteicas, estudo filogenético, busca e anotação de sequências genômicas, além de integração com aprendizado de máquina e estatísticas (Zulkower; Rosser, 2020).

A biblioteca Biopython oferece uma interface simples e intuitiva para trabalhar com sequências biológicas. É possível ler e gravar sequências em diversos formatos, realizar operações como transcrição, tradução e complementação reversa, bem como acessar informações específicas de uma sequência.

No contexto dos alinhamentos de sequências, *Biopython* disponibiliza algoritmos populares como BLAST, ClustalW e Muscle, permitindo a comparação de sequências e a identificação de regiões conservadas. Também é possível manipular e visualizar alinhamentos de forma eficiente. Para análise de estruturas proteicas, a Biopython oferece recursos para obter dados do Banco de Dados de Proteínas (PDB), calcular propriedades estruturais, realizar superposições de estruturas e visualizá-las usando ferramentas externas.

Biopython é amplamente utilizado na genômica, com recursos para análise de dados genômicos, extração de características e anotações de sequências, busca e recuperação de informações do GenBank, e estudo de variações genéticas (McBroome; Turakhia; Corbett-Detig, 2022).

2.2.3.2 Genbank

O NCBI (National Center for Biotechnology Information) é uma instituição renomada e fornecem recursos essenciais para a comunidade científica e a indústria biomédica. Ele foi criado com o objetivo de coletar, armazenar e fornecer acesso a uma vasta quantidade de informações biológicas e genômicas. O NCBI desempenha um papel crucial na pesquisa em biologia molecular, genômica, biotecnologia e medicina, além de ser uma referência fundamental para o estudo da evolução e da diversidade genética. Uma das principais bases de dados mantidas pelo NCBI é o Genbank.

O Genbank é um repositório abrangente de sequências de DNA (ácido desoxirribonucleico), RNA (ácido ribonucleico) e proteínas. Ele contém informações genéticas de uma ampla variedade de organismos, incluindo vírus, bactérias, plantas e animais. O Genbank permite o acesso e a pesquisa de sequências genômicas, bem como o compartilhamento de dados entre cientistas de todo o mundo. Essa base de dados tem sido fundamental para o avanço da pesquisa em genômica, possibilitando a análise comparativa de genes, o estudo da expressão gênica e a identificação de novos alvos terapêuticos (Benson *et al.*, 2012).

A disponibilidade do Genbank como um repositório abrangente de sequências genéticas tem impulsionado significativamente a pesquisa científica em várias áreas. Através dessa valiosa fonte de dados, os pesquisadores têm acesso a uma vasta gama de informações genômicas, o que lhes permite realizar estudos comparativos de genes entre diferentes espécies. O Genbank revolucionou a taxonomia e se tornou um repositório fundamental

para compreender a evolução e a diversidade genética, bem como para investigar os mecanismos moleculares subjacentes a diferentes características biológicas (Rapini, 2004). Além disso, o Genbank em 2020 contava com 1,6 bilhão de sequências de nucleotídeos de 450.000 espécies formalmente descritas (Sayers *et al.*, 2020). Já janeiro de 2022, seu repositório passou a possuir aproximadamente 2,5 bilhões de sequências de nucleotídeos disponíveis publicamente (Sayers *et al.*, 2020).

2.2.3.3 Celery

O Celery é uma biblioteca de código aberto projetada para a gestão eficiente de requisições de software em sistemas distribuídos (Eichler; Şahin; Gurevych, 2019). Baseado no paradigma de computação assíncrona, ele adota uma arquitetura cliente-servidor que permite o processamento ágil e otimizado de tarefas assíncronas e a distribuição de carga em diferentes recursos computacionais. A gestão de requisições em sistemas distribuídos é um desafio complexo enfrentado por desenvolvedores e pesquisadores devido à natureza assíncrona das aplicações modernas e à necessidade de escalabilidade e eficiência (Eichler; Şahin; Gurevych, 2019). O *Celery* surge como uma solução efetiva, fornecendo um *framework* flexível e altamente eficiente para a gestão de tarefas assíncronas.

A arquitetura do Celery consiste em três componentes principais: cliente, workers e broker. O cliente envia tarefas assíncronas para processamento, enquanto os *workers* processam e executam essas tarefas. O *broker* atua como intermediário, gerenciando a fila de tarefas e garantindo a comunicação assíncrona. Essa arquitetura permite a distribuição eficiente das tarefas e o controle preciso de sua execução e escalabilidade. Em bioinformática o *Celery* vem sendo amplamente aplicado em segundo plano dos sistemas nas requisições de processamento e análise dos dados, porque permite que requisições de usuários que demandam alto poder de processamento e tempo de execução possam ser gerenciadas corretamente e realizadas de maneira eficaz (Gomes; Queiroz; Ferreira, 2020).

2.2.3.4 RabbitMQ

O *RabbitMQ* é um sistema de mensagens assíncronas que implementa o protocolo AMQP, fornecendo uma arquitetura robusta para o envio e recebimento de mensagens entre diferentes componentes de um sistema distribuído (Nugroho; Kusumawardani *et al.*, 2020). Sua arquitetura é composta por produtores, consumidores e uma camada intermediária chamada de *broker*, responsável pelo roteamento e armazenamento das mensagens. Essa estrutura permite uma comunicação assíncrona eficiente e flexível, garantindo a escalabilidade e confiabilidade necessárias para lidar com o processamento distribuído de dados genômicos.

O *RabbitMQ* oferece uma série de recursos e benefícios relevantes para a bioinformática. A escalabilidade é um dos principais aspectos, permitindo que o sistema seja

dimensionado de acordo com a demanda, facilitando a distribuição de tarefas e a integração de novos nós de processamento (Nugroho; Kusumawardani *et al.*, 2020). Além disso, o *RabbitMQ* possui mecanismos robustos de tolerância a falhas, garantindo a confiabilidade das comunicações em ambientes distribuídos e reduzindo a possibilidade de perda de dados genômicos importantes.

3 TRABALHOS CORRELATOS

Neste capítulo serão apresentados alguns trabalhos do levantamento referencial teórico, consiste em tecnologias e ferramentas implementados utilizando Inteligência Artificial e *Deep Learning*.

3.1 Estado da Arte

O nosso OKR 1 (revisão sistemática), foi realizada utilizando a ferramenta Parsifal, na versão V2.2. O Parsifal é uma ferramenta *online* que auxilia os pesquisadores na condução de revisões sistemáticas da literatura.

- Palavras-chave e Sinônimos

Foram selecionadas cinco palavras-chave principais: *Deep Learning*, *Genes*, *Prediction*, *Method* e *Genome mining*. Cada uma dessas palavras possui um significado específico e é relevante para a análise do tema em questão. Além disso, para aumentar a efetividade da busca, foram incluídos alguns sinônimos, tais como *Proteins*, *Tool*, *Algorithm* e *Architecture*. Demonstradas na Tabela 1.

Tabela 1 – Palavras-chave e sinônimos utilizados nas buscas

| Tabela de Palavras-chaves | Sinônimos |
|---------------------------|--------------------------------------|
| <i>Deep Learning</i> | - |
| <i>Genes</i> | <i>Proteins</i> |
| <i>Prediction</i> | - |
| <i>Method</i> | <i>Tool, Algorithm, Architecture</i> |
| <i>Genome mining</i> | - |

Fonte: Produzido pelo autor (2024)

- Critérios de Inclusão e Exclusão.

Foram definidos 5 (cinco) critérios para a seleção de artigos, sendo 4 (quatro) de inclusão e 1 (um) de exclusão. O I1 - Para restringir estudos que mencionem processos de identificação de genes. I2 - Para retornar artigos que mencionassem mineração em genomas. I3 - Para aceitar artigos e estudos indicados pelo orientador. E1 - Para excluir artigos que não se enquadrassem no tema da revisão. Como demonstra a Tabela 2.

Tabela 2 – Critérios de inclusão e exclusão dos artigos

| Identificador | Critérios |
|---------------|--|
| I2 | Menciona a identificação de genes |
| I3 | Menciona mineração de genoma |
| I4 | Recomendado pelo orientador |
| E1 | Menciona o método computacional (<i>Deep Learning</i>) |

Fonte: Produzido pelo autor (2024)

Realizamos uma pesquisa em três bases de dados renomadas para encontrar artigos relevantes: *PubMed*, *Science Direct* e *SpringerLink*. Na *PubMed*, encontramos inicialmente 38 resultados, que foram reduzidos para 5 (cinco) artigos após a aplicação dos critérios de seleção. Na *Science Direct*, encontramos um total de 57 artigos, dos quais, 8 (Oito) se encaixaram nos critérios de seleção. Na base *SpringerLink*, obtivemos 45 (Quarenta e cinco) artigos e, após a seleção, restaram 4 artigos. Além disso, o orientador também indicou 4 (Quatro) artigos adicionais, complementando nossa pesquisa.

Dois grupos de artigos foram separados, um grupo de artigos contém implementação de uma ferramenta, método ou aplicação e os artigos sem implementação que serão utilizados para compreensão dos técnicas de Aprendizado profundo, suas aplicações no meio genômico/ genético, conceitos e limitações, além de embasar estudos o nosso trabalho.

Os resultados com implementação foram 15 artigos que apresentaram métodos, aplicações e ferramentas. Demonstrados na Tabela 3:

Tabela 3 – Tabela de Ferramentas e aplicações

| Aplicação desenvolvida | Ferramentas e Bibliotecas utilizadas |
|---|---|
| DeepBGC (Hannigan <i>et al.</i> , 2019) | <ul style="list-style-type: none"> • Python: Implementação do framework em Python • Keras e TensorFlow: Utilizado para implementar a rede neural profunda |
| DeepEc (Ryu; Kim; Lee, 2019) | <ul style="list-style-type: none"> • DIAMOND: Utilizado para lidar com grandes volumes de dados de sequências genômicas. |
| DeepTFactor (Kim <i>et al.</i> , 2021) | <ul style="list-style-type: none"> • Python-Implementação do framework em Python. • SwissProt- Conjunto de dados. |

| Aplicação desenvolvida | Ferramentas e Bibliotecas utilizadas |
|--|--|
| DeepLFT (Kong <i>et al.</i> , 2021) | <ul style="list-style-type: none"> • AutoEncoder: Utilizado para construir a rede neural duas camadas profundas • Python: Implementação do framework em Python. • TensorFlow: Utilizado para implementar a rede neural profunda. |
| Deep-LC (Li <i>et al.</i> , 2022a) | <ul style="list-style-type: none"> • Python: Implementação do framework em Python. • TensorFlow: Utilizado para implementar a rede neural profunda. • Gene Expression Omnibus (GEO): Conjunto de dados. • Gene Ontology (GO): Conjunto de dados • Kyoto Encyclopedia of Genes and Genomes (KEGG): Conjunto de dados |
| MU-PseUDep (Khan <i>et al.</i> , 2020) | <ul style="list-style-type: none"> • Python: Implementação do framework em Python. • TensorFlow: Utilizado para implementar a rede neural profunda. • Scikit-learn: Utilizado para pré-processar os dados. • RMBase v2.0: Conjunto de dados. |
| Entendimento dos métodos (Monaco <i>et al.</i> , 2021) | <ul style="list-style-type: none"> • TensorFlow: Utilizado para implementar a rede neural. • PyTorch: Utilizado para implementar a rede neural. |

| Aplicação desenvolvida | Ferramentas e Bibliotecas utilizadas |
|---|--|
| RegCNN (Yang; Yang; Tu, 2022) | <ul style="list-style-type: none"> • Python: Implementação do framework em Python. • TensorFlow: Utilizado para implementar a rede neural. • MARC: (Model-based Analysis of ChIP-Seq). |
| ProtInteract (Soleymani <i>et al.</i> , 2023) | <ul style="list-style-type: none"> • Python: Implementação em Python. • TensorFlow: Utilizado para implementar a rede neural • Scikit-learn: Utilizado para pré-processar os dados. • Proteína Data Bank (PDB): Conjunto de dados. |
| CDRGI (Al-Obeidat <i>et al.</i> , 2022) | <ul style="list-style-type: none"> • R: Utilizado para realizar a análise exploratória dos dados de expressão gênica • Gene Expression Omnibus (GEO): GEO é utilizado para coletar os dados de expressão gênica • Random Forest: utilizado para avaliar a performance do método proposto de identificação de genes relevantes • SVM (Support Vector Machine): SVM é utilizado para classificar os dados de expressão gênica. |

| Aplicação desenvolvida | Ferramentas e Bibliotecas utilizadas |
|---|--|
| GoldAI Sachs (Serrano, 2020) | <ul style="list-style-type: none">• Python: Implementação em Python• Keras e TensorFlow: Utilizados como backend para implementar a rede neural.• Scikit-learn: Utilizado para pré-processar os dados.• NumPy: Utilizado para realizar operações matemáticas.• Pandas: Utilizado para pré-processar os dados e manipular os dataframes.• Algoritmos genéticos. |
| Método Implementado em Python (Uzma <i>et al.</i> , 2022) | <ul style="list-style-type: none">• Keras e TensorFlow: Utilizados o Keras com TensorFlow como backend para implementar a rede neural autoencode.• Scikit-learn-Algoritmo K-means: utilizado para agrupar os genes.• Pandas: Utilizado para pré-processar os dados de expressão gênica.• NumPy: Utilizado para realizar operações matemáticas• Matplotlib: Utilizado para visualizar os resultados |

| Aplicação desenvolvida | Ferramentas e Bibliotecas utilizadas |
|---|--|
| Modelo SVM (Ghosh; Ghosh, 2021) | <ul style="list-style-type: none"> • R: Utilizado para realizar análises exploratórias dos dados de expressão gênica antes de aplicar o modelo proposto. • Fuzzy Toolbox: Utilizada para implementar a abordagem de multigranulação tipo-2. • LibSVM: Utilizado para treinar e testar o modelo SVM proposto. • Gene Expression Omnibus (GEO): Conjunto de dados de expressão gênica • MATLAB: Utilizado para implementar o modelo proposto e realizar as análises de dados. |
| DeepTMHMM (Hallgren <i>et al.</i> , 2022) | <ul style="list-style-type: none"> • Python: Software construído em python com código disponível para utilização. • Pode ser instalado utilizando o comando pip. |
| MebiPred (Aptekmann <i>et al.</i> , 2022) | <ul style="list-style-type: none"> • Python: Software construído em python com código disponível para utilização. • Keras e TensorFlow: Utilizados o Keras com TensorFlow como backend para implementar a rede neural autoencode. • Pode ser instalado utilizando o comando pip direto ou baixando o pacote como a extensão .tar.gz. |

Fonte: Produzido pelo autor (2024)

Dentre os softwares e ferramentas encontrados, dois se destacaram por realizar funções importantes no nosso trabalho: DeepTMHMM e MebiPred.

3.1.1 *Mebipred (MyMetal)*

O software Metal Binding Predictor (*Mebipred*) é uma ferramenta avançada para processamento de dados de sequências de proteínas, empregando técnicas de aprendizado de máquina para identificar o potencial de ligação a metais nessas sequências. Sua saída consiste em previsões binárias, indicando se uma determinada sequência de proteína possui potencial para se ligar a metais, com base nas características das sequências analisadas e no modelo de aprendizado de máquina treinado.

Diferentemente dos métodos tradicionais, o *Mebipred* não se baseia em alinhamentos de sequências. Em vez disso, este método utiliza inteligência artificial através do aprendizado do máquina em suas análises. Para otimizar o processamento, os dados são analisados utilizando multiprocessamento, onde são divididos, resultando em análises mais rápidas e eficientes. Utilizando recursos derivados das próprias sequências, o *Mebipred* é capaz de realizar previsões precisas, mesmo a partir de trechos curtos de sequências, como leituras de sequenciamento traduzidas. Essa capacidade o torna útil para a anotação de requisitos metálicos em amostras metagenômicas, como microbiomas de diversos ambientes.

Para criar o *Mebipred*, foram utilizadas propriedades específicas para treinar um perceptron de várias camadas utilizando a implementação do Keras no Tensorflow. Os programadores escolheram um arranjo sequencial com o otimizador RMSprop e uma taxa de aprendizado de 0,000005, ajustando essa taxa ao longo do processo de treinamento para minimizar a perda. Cada iteração do modelo compreendeu 1000 épocas, com camadas compostas por 219 neurônios, incluindo duas camadas ocultas ativadas pela função ReLU e com uma taxa de dropout de 0,2, culminando em uma camada de saída ativada pela função sigmóide.

Adicionalmente, para validar a efetividade do modelo na detecção de proteínas ligantes de metais, utilizou-se a técnica de validação cruzada de 10 folds. O conjunto de dados foi dividido em 10 grupos equitativos, alternando entre eles para treinar e testar o modelo em cada iteração. Esse procedimento permitiu uma avaliação ampla do desempenho do modelo, sendo o modelo final construído a partir de todas as sequências positivas e negativas disponíveis.

O *Mebipred* demonstrou eficácia na previsão do potencial de ligação a metais em proteínas, alcançando uma área sob a curva de precisão/recall de 0,91. Isso evidencia sua alta precisão na identificação de proteínas com potencial de ligação a metais. Além disso, ao ser avaliado em conjuntos de proteínas não previamente observados, o *Mebipred* apresentou um F1 máximo de 0,74 e uma AUPRC de 0,72, indicando sua capacidade de

generalização e consistência de desempenho em diferentes conjuntos de dados (Aptekmann *et al.*, 2022).

o *Mebipred* se destaca por sua capacidade de utilizar multiprocessamento para analisar arquivos ou sequências submetidas a ele de maneira rápida e eficiente. Esta ferramenta consegue identificar quais proteínas se ligam a 10 elementos diferentes, sendo 8 deles micronutrientes essenciais e 2 elementos essenciais à saúde humana. A utilização do MebiPred nos permite não apenas acelerar o processo de análise, mas também obter resultados precisos sobre as interações proteicas com micronutrientes, contribuindo significativamente para a compreensão das funções biológicas e potencial aplicação em biotecnologia.

Este software processa arquivos no formato FASTA, comumente utilizado para representar sequências de proteínas. Os arquivos FASTA contêm informações sobre a sequência de aminoácidos de proteínas específicas, e o *Mebipred* é capaz de analisar esses dados para prever o potencial de ligação a metais nas proteínas listadas. Com uma precisão de mais de 80% na identificação de proteínas que se ligam a ligantes contendo íons metálicos, e com capacidade de reconhecer a identidade específica de 11 íons metálicos, incluindo Cálcio (Ca), Cobalto(Co), Cobre(Cu), Ferro(Fe), Potássio(K), Manganês(Mg), Magnésio(Mn), Sódio(Na), Níquel (Ni), Zinco (Zn) o *Mebipred*, destaca-se como uma ferramenta valiosa para análise de proteínas em diversos contextos acadêmicos e industriais. O Quadro 2 demonstrado os elementos em que o MebiPred identifica em suas predições, em relação a elementos que os cogumelos acumulam.

Quadro 2 – Acúmulo de elementos inorgânicos em cogumelos e predição de proteínas que se ligam a íons metálicos

| Elementos acumulados por cogumelos | Elementos que o MebiPred prediz |
|------------------------------------|---------------------------------|
| Arsênio (As) | - |
| Boro (B) | - |
| Bromo (Br) | - |
| Cádmio (Cd) | - |
| Cálcio (Ca) | Cálcio (Ca) |
| Césio (Cs) | - |
| Cloro (Cl) | - |
| Cobalto (Co) | Cobalto (Co) |
| Cobre (Cu) | Cobre (Cu) |
| Cromo (Cr) | - |
| Ferro (Fe) | Ferro (Fe) |
| Flúor (F) | - |
| Fósforo (P) | - |
| Iodo (I) | - |
| Lítio (Li) | - |
| Magnésio (Mg) | Magnésio (Mg) |
| Manganês (Mn) | Manganês (Mn) |
| Mercúrio (Hg) | - |
| Molibdênio (Mo) | - |
| Níquel (Ni) | Níquel (Ni) |
| Ouro (Au) | - |
| Potássio (K) | Potássio (K) |
| Prata (Ag) | - |
| Rubídio (Rb) | - |
| Selênio (Se) | - |
| Sódio (Na) | Sódio (Na) |
| Vanádio (V) | - |
| Zinco (Zn) | Zinco (Zn) |

Fonte: Retirada de (Falandysz; Borovička, 2013; Aptekmann *et al.*, 2022)

3.1.2 DeepTMHMM

O *DeepTMHMM* é um método avançado de aprendizagem profunda utilizado para a predição da topologia de proteínas transmembrana. Esse método se destaca como um dos mais completos e eficientes para a determinação da estrutura de proteínas com regiões transmembrana alfa-helicoidais e beta-barril (Hallgren *et al.*, 2022). Sua abordagem consiste em codificar a sequência de aminoácidos primários por meio de um modelo de linguagem pré-treinado e, posteriormente, decodificar a topologia usando um modelo de espaço de estado. Essa combinação permite a obtenção de previsões precisas e sem precedentes sobre a topologia e o tipo de proteína transmembrana.

O *DeepTMHMM* é capaz de escanear proteomas completos, fazendo previsões da

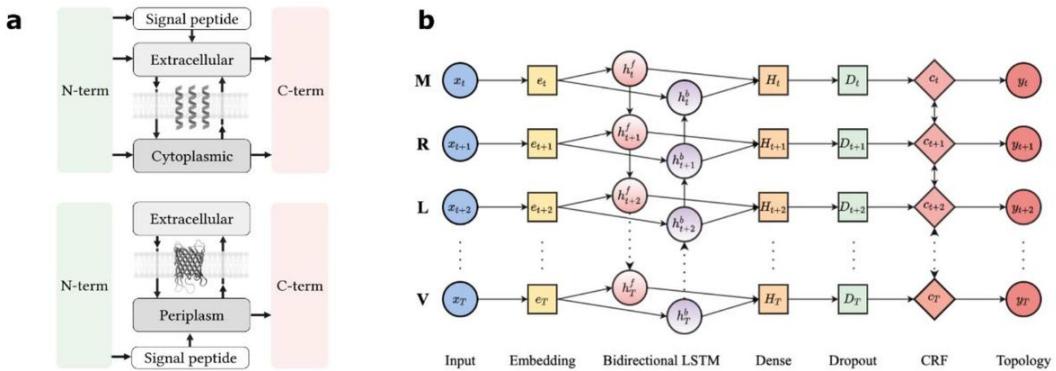
topologia, identificando e caracterizando ambas as classes de proteínas transmembrana, tanto as que possuem estruturas de hélice alfa, quanto as com estrutura de barril beta. Sua aplicação tem sido extremamente valiosa para a comunidade científica, proporcionando uma compreensão mais aprofundada sobre a estrutura e a função dessas proteínas essenciais. Além disso, sua alta precisão na predição da topologia permite o desenvolvimento de terapias direcionadas a proteínas transmembrana, que são alvos fundamentais para o desenvolvimento de fármacos.

O DeepTMHMM será utilizado para identificar as proteínas transmembrana existentes no genoma a ser analisado. Este método foi construído na linguagem Python, o que facilita sua integração com outras ferramentas e pipelines de bioinformática. O DeepTMHMM foi treinado utilizando cinco tipos de proteínas, permitindo a identificação precisa de proteínas transmembrana helicoidais sem peptídeo sinal (alfa TM), proteínas transmembrana com peptídeo sinal (SP + alfa TM), proteínas transmembrana de barril beta, proteínas globulares com peptídeo sinal (SP + Globular) e proteínas globulares sem peptídeo sinal (Globular). A precisão e eficiência do DeepTMHMM na predição destaca-se em comparação a outros, isso tornou-o uma escolha ideal para nossos objetivos.

Com a utilização do *DeepTMHMM*, é possível obter de proteínas transmembrana com uma acurácia sem precedentes. Essa ferramenta avançada e eficiente tem o potencial de impulsionar significativamente a pesquisa em biologia molecular, especialmente no campo do desenvolvimento de fármacos. Ao fornecer informações precisas sobre a topologia das proteínas transmembrana, o *DeepTMHMM* abre novas possibilidades para a identificação de alvos terapêuticos e a concepção de estratégias terapêuticas mais direcionadas e eficazes.

A arquitetura do *DeepTMHMM*, que pode ser vista na Figura 8, combina técnicas de aprendizado profundo (Bochie *et al.*, 2020), processamento de sequência e modelagem de dependência para alcançar resultados precisos na tarefa de previsão de topologia de proteínas. Sua capacidade de capturar informações estruturais e evolutivas implícitas na sequência de aminoácidos, juntamente com o uso de mecanismos de atenção e modelos de linguagem pré-treinados, contribui para seu desempenho superior. A aplicação dessa arquitetura pode ter impactos significativos na compreensão e no desenvolvimento de terapias direcionadas a proteínas transmembranares e em outras áreas da biologia molecular.

Figura 8 – Arquitetura do DeepTMHMM



Fonte: Retirada de (Hallgren *et al.*, 2022)

A arquitetura do *DeepTMHMM* combina técnicas de aprendizado profundo, processamento de sequência e modelagem de dependência para alcançar resultados precisos na tarefa de previsão de topologia de proteínas. Sua capacidade de capturar informações estruturais e evolutivas implícitas na sequência de aminoácidos, juntamente com o uso de mecanismos de atenção e modelos de linguagem pré-treinados, contribui para seu desempenho superior (Hallgren *et al.*, 2022).

O decodificador do modelo é composto por uma LSTM bidirecional (Sampaio, 2022) que recebe as representações do codificador como entrada. Ele utiliza mecanismos de atenção para capturar as relações de dependência entre os resíduos de aminoácidos e gerar uma sequência de rótulos por resíduo correspondente. A atenção permite que o decodificador se concentre nas partes relevantes da sequência de entrada durante a geração dos rótulos, melhorando a precisão e coerência do resultado final (Hallgren *et al.*, 2022).

Durante o treinamento, o modelo é alimentado com pares de sequências de proteínas e suas respectivas sequências de rótulos por resíduo. A função de perda é calculada comparando os rótulos gerados pelo modelo com os rótulos reais. O algoritmo de retropropagação é utilizado para ajustar os parâmetros do modelo e minimizar a diferença entre as previsões e os rótulos verdadeiros.

Uma vez treinado, o *DeepTMHMM* é capaz de prever a topologia de uma proteína desconhecida com base em sua sequência de aminoácidos. Isso é extremamente útil, pois a topologia da proteína desempenha um papel fundamental em sua função e interações com outras moléculas. O conhecimento da topologia permite inferir informações sobre como a proteína se dobra e se posiciona em relação às membranas celulares, por exemplo (Hallgren *et al.*, 2022).

Ao executar o software, ele identifica no arquivo submetido quantas sequências de proteínas existem no arquivo .fasta. Em seguida, ele inicia a primeira fase, "Loading

"transformer model", que consiste apenas em carregar o modelo. Em seguida, passa para a segunda etapa, "Generating embeddings for sequences", que se refere ao processo de representar dados em um espaço vetorial, chamado de espaço de incorporação. Isso é feito convertendo os dados brutos em vetores numéricos de características, chamados *embeddings*, que capturam informações relevantes sobre os dados de entrada. Na terceira etapa, "Predicting topologies for sequences in batches of 1", o software prediz a topologia de cada proteína analisada. A quarta e última etapa, "Generating output", é onde os resultados são produzidos, todo esse fluxo pode ser visto na Figura 9.

Figura 9 – Exemplo do processamento do DeepTMHMM

```
[2024-06-02 17:11:59,694: WARNING/ForkPoolWorker-8] Running DeepTMHMM on 4 sequences...
Step 1/4 | Loading transformer model...
[2024-06-02 17:12:13,652: WARNING/ForkPoolWorker-8]
Step 2/4 | Generating embeddings for sequences...
Generating embeddings:  0% 0/4 [00:00<?, ?seq/s]8]
Generating embeddings: 25% 1/4 [00:18<00:54, 18.09s/seq]
Generating embeddings: 50% 2/4 [00:37<00:37, 18.68s/seq]
Generating embeddings: 75% 3/4 [00:45<00:13, 13.97s/seq]
Generating embeddings: 100% 4/4 [00:50<00:00, 10.33s/seq]
Generating embeddings: 100% 4/4 [00:50<00:00, 12.57s/seq]
[2024-06-02 17:13:04,471: WARNING/ForkPoolWorker-8]
Step 3/4 | Predicting topologies for sequences in batches of 1...
Topology prediction:  0% 0/4 [00:00<?, ?seq/s]
Topology prediction: 25% 1/4 [00:10<00:32, 10.73s/seq]
Topology prediction: 50% 2/4 [00:21<00:20, 10.48s/seq]
Topology prediction: 75% 3/4 [00:26<00:08,  8.11s/seq]
Topology prediction: 100% 4/4 [00:29<00:00,  6.05s/seq]
Topology prediction: 100% 4/4 [00:29<00:00,  7.30s/seq]

Step 4/4 | Generating output...
```

Fonte: Produzida pelo autor

O referencial teórico apresentado e os trabalhos correlatos dão suporte para solução do problema de pesquisa e para atingir os objetivos desta pesquisa, conforme metodologia apresentada no próximo capítulo.

4 METODOLOGIA

Este capítulo seleciona a modalidade de pesquisa adequada ao objeto de pesquisa, identificando o tipo de pesquisa quanto à sua natureza, seus objetivos e seus procedimentos. A pesquisa é de natureza exploratória, com objetivos aplicados e procedimentos baseados na Design Science Research (DSR).

4.1 Metodologia científica

Este estudo apresenta um pipeline para prospecção de cogumelos com capacidade de acumular micronutrientes, ou seja, trata-se de uma pesquisa aplicada. O objetivo principal dessa abordagem é facilitar as pesquisas relacionadas à bioacumulação de elementos essenciais e micronutrientes em cogumelos, fornecendo uma visão abrangente sobre quais genes estão envolvidos nesse processo.

Neste trabalho, quanto aos procedimentos, foi adotada a metodologia de pesquisa conhecida como *Design Science Research (DSR)*. O DSR é uma abordagem amplamente reconhecida que visa criar, desenvolver e avaliar soluções práticas para problemas complexos e relevantes no campo da ciência da computação e outras disciplinas relacionadas (Brocke; Hevner; Maedche, 2020). Essa abordagem coloca ênfase na criação de artefatos, como sistemas, modelos, processos ou *frameworks*, que podem ser usados para resolver problemas reais. A explicação da aplicação do DSR na nossa pesquisa pode ser observada na Figura 10.

O DSR segue um processo iterativo, composto por várias fases distintas. Inicialmente, ocorre a identificação e compreensão do problema, onde são analisados o contexto, as necessidades e os requisitos dos usuários envolvidos. Com base nessa análise, são definidos os objetivos e as especificações do artefato a ser desenvolvido.

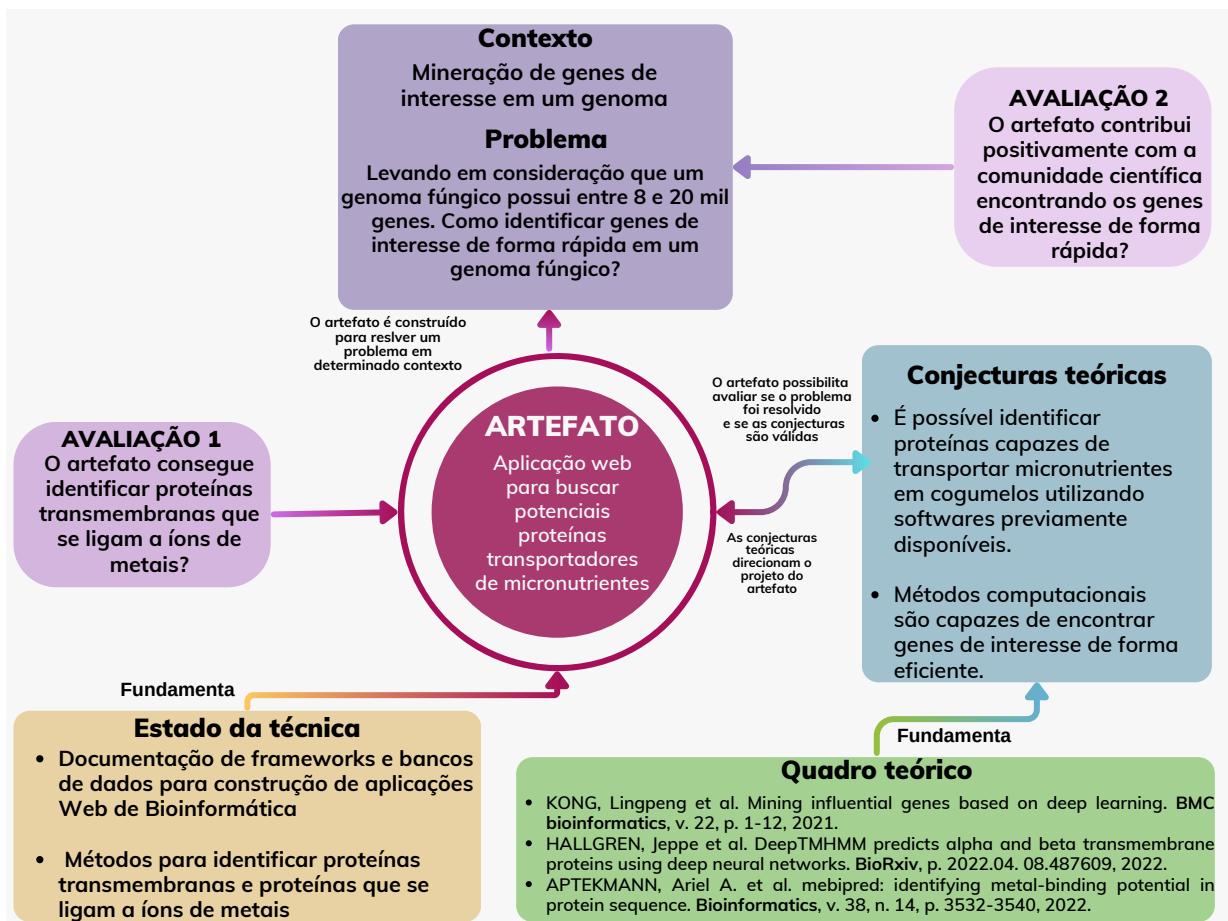
Em seguida, entra-se na fase de design, em que são propostas soluções conceituais e criativas para abordar o problema identificado. Essas soluções são fundamentadas em teorias, princípios e melhores práticas existentes na área de estudo. Durante o processo de design, podem ser desenvolvidos protótipos ou modelos iniciais do artefato na finalidade de avaliar e refinar.

Após o design, segue-se a fase de construção, na qual o artefato é implementado de acordo com as especificações definidas na fase anterior. Essa implementação pode envolver o desenvolvimento de software, a criação de hardware, a construção de estruturas físicas ou outras atividades relevantes, dependendo da natureza do problema e do artefato proposto.

Uma vez construído, o artefato passa por uma fase de avaliação, na qual é testado e avaliado quanto à sua eficácia, eficiência e adequação aos requisitos estabelecidos. Os resultados da avaliação são utilizados para aprimorar o artefato, se necessário, e para fornecer *insights* valiosos para a compreensão do problema e o avanço do conhecimento científico na área.

Além disso, o DSR enfatiza a reflexão crítica sobre o processo de pesquisa e o aprendizado obtido ao longo do caminho. Isso envolve a análise dos resultados, a discussão das limitações do estudo e a identificação de oportunidades de pesquisa futuras.

Figura 10 – *Design Science Research (DSR)*



Fonte: Produzido pelo autor (2024)

4.2 Metodologia de Desenvolvimento

Como metodologia de desenvolvimento da presente pesquisa foi usada a gestão ágil, que tem mostrado uma abordagem eficaz não apenas no desenvolvimento de projetos de software, mas também na condução de pesquisas (Nagai; Sbragia, 2023). Essa metodologia proporciona uma maneira flexível e adaptável de realizar pesquisas, permitindo uma maior agilidade e eficiência no processo. A gestão ágil pode ajudar os pesquisadores a lidar com a

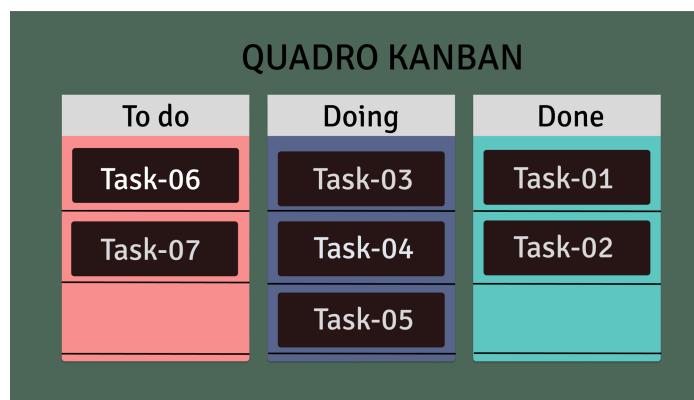
incerteza inerente ao processo de pesquisa, permitindo que eles respondam rapidamente às mudanças e ajustem suas estratégias conforme necessário. Esse projeto de pesquisa utiliza um *framework* de gestão ágil baseado em *Objectives and Key Results (OKR)*, *Scrum* e *Kanban* demonstrado na Figura 11.

Os OKRs são uma forma de realizar o planejamento estratégico partindo dos objetivos específicos do projeto e de como medi-los para determinar se os resultados esperados foram obtidos com sucesso ou não. Com os OKRs, define-se qual será o rumo que guiará o projeto em um determinado período de tempo (Cardoso, 2020).

O *Scrum* é utilizado para realizar o trabalho operacional de forma inteligente e bem planejada, de maneira a atingir os OKRs propostos. O trabalho operacional geralmente corresponde à implementação de métodos computacionais e ferramentas de software. Esse trabalho operacional é dividido em ciclos chamados *Sprints*, onde as tarefas do seu planejamento são executadas por ordem de valor, visando gerar entregas rápidas e relevantes (Hossain; Babar; Paik, 2009). No *Scrum*, os papéis são bem distribuídos e alinhados à multidisciplinaridade de cada um dos membros do time. Destacando-se as reuniões semanais que o grupo realiza para gestão e controle dos projetos.

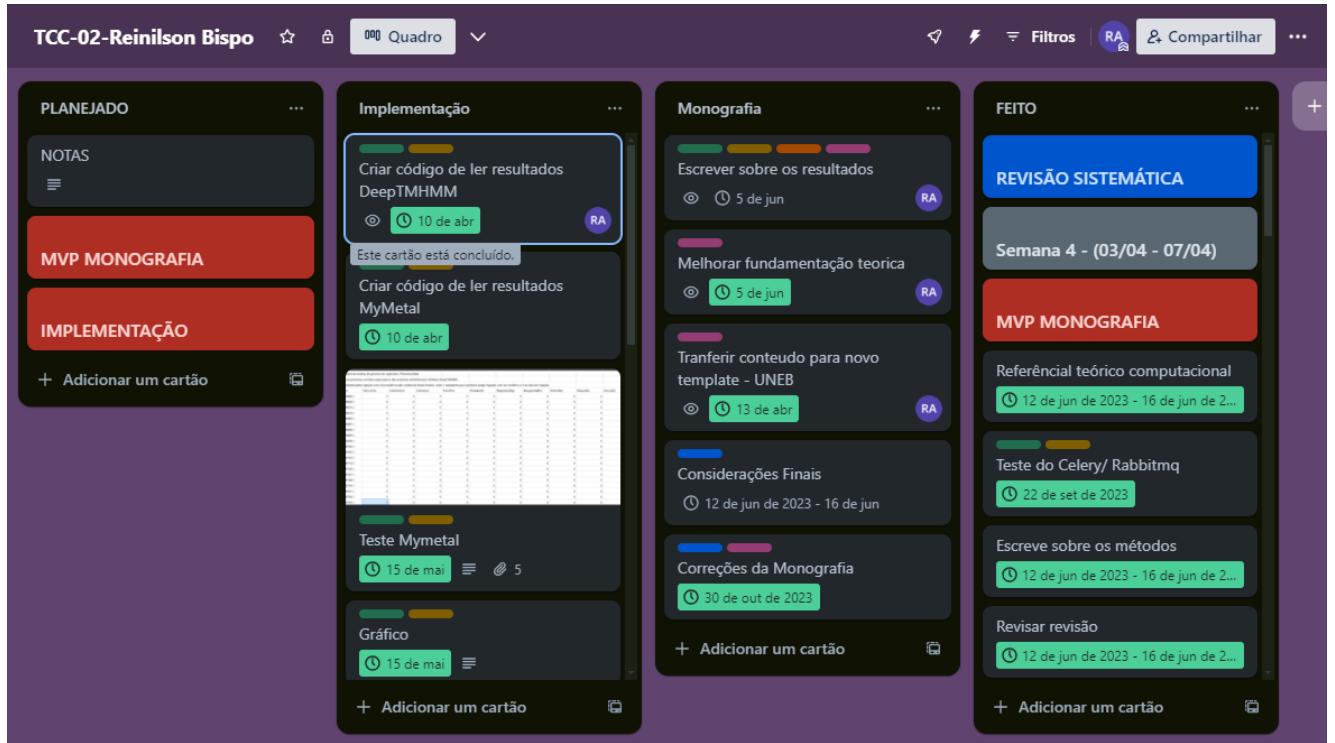
Por fim, o Kanban é uma metodologia japonesa para gestão e controle de atividades que adaptamos para estar alinhada ao Scrum (Ahmad; Markkula; Oivo, 2013). Com auxílio da ferramenta on-line *Trello* demonstrado no Figura 12, escrevemos todas as etapas e tarefas necessárias, organizadas em *Sprints*, para que todos os reforços serão focadas apenas nessas tarefas e manter o foco. Cada tarefa das *Sprints* segue o fluxo de trabalho esse fluxo é demostrado na Figura 11, que é dividido em três: “To do”, que significa “A fazer”, “Doing”, que é “O que está sendo feito” e “Done”, que significa “O que está finalizado”.

Figura 11 – Quadro Kanban.



Fonte: Produzido pelo autor (2024)

Figura 12 – Interface do Trello.



Fonte: Produzido pelo autor (2024)

Os resultados são apresentados no capítulo subsequente.

5 RESULTADOS

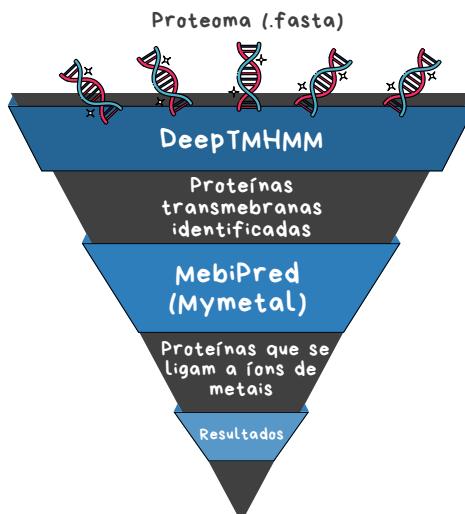
Neste capítulo são apresentados os resultados obtidos durante a pesquisa conforme OKRs definidos e organizados em duas seções:

- Método computacional(Aplicação Web)
- Validação Quantitativa (Tempo) e Qualitativa (Funcionalidade)

5.1 Método Computacional(Pipeline)

O método implementa um procedimento computacional que minera um arquivo de proteoma de um fungo para buscar e identificar proteínas transmembrana que transportam micronutrientes e outros elementos essenciais para nutrição humana. O pipeline possui dois módulos principais: o software DeepTMHMM, que realiza a identificação das proteínas transmembrana, e o Mebipred, que identifica as proteínas que se ligam a íons de metais. Além disso, dois filtros realizam etapas intermediárias, filtrando a saída de cada um dos softwares. A pipeline implementada pode ser vista na Figura 13

Figura 13 – Representação gráfica das etapas do pipeline



Fonte: Produzido pelo autor (2024).

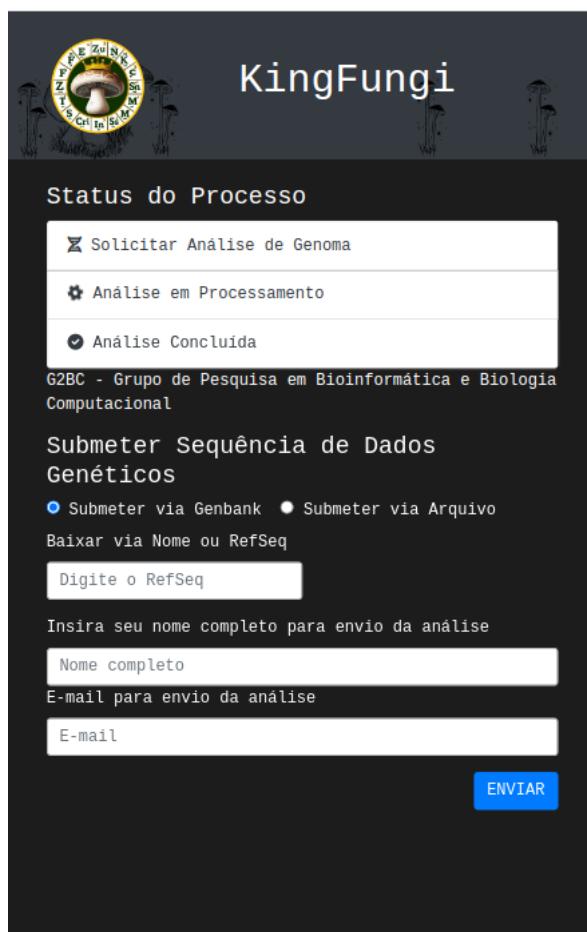
5.1.1 Arquitetura da aplicação

Para viabilizar o funcionamento contínuo do pipeline, desenvolvemos uma aplicação web chamada “KingFungi”. Optamos pelo desenvolvimento web devido aos diversos benefícios que essa tecnologia oferece. Primeiramente, proporciona escalabilidade, facilitando o

acesso e uso por usuários com diferentes níveis de conhecimento computacional. Além disso, a tecnologia web melhora significativamente a realização das análises, uma vez que não demanda recursos computacionais intensivos. As análises podem ser facilmente iniciadas através de um navegador em dispositivos móveis (Figura 14) ou *desktops* (Figura 15).

A aplicação proposta representa uma abordagem prática para a análise de genomas, oferecendo aos usuários uma interface acessível e simples para submissão de genomas, essa interface pode ser vista na Figura 15. Sua funcionalidade permite aos usuários escolher entre três métodos de submissão: inserir o nome científico do organismo a partir do banco de dados Genbank (NCBI), fornecer o Reference Sequence (RefSeq) como identificador do genoma, ou submeter o arquivo .FASTA. Essa flexibilidade na entrada de dados busca simplificar o processo de submissão e tornar a plataforma mais flexível para os usuários.

Figura 14 – Front-end KingFungi Mobile



Fonte: Produzido pelo autor (2024).

Figura 15 – Front-end KingFungi

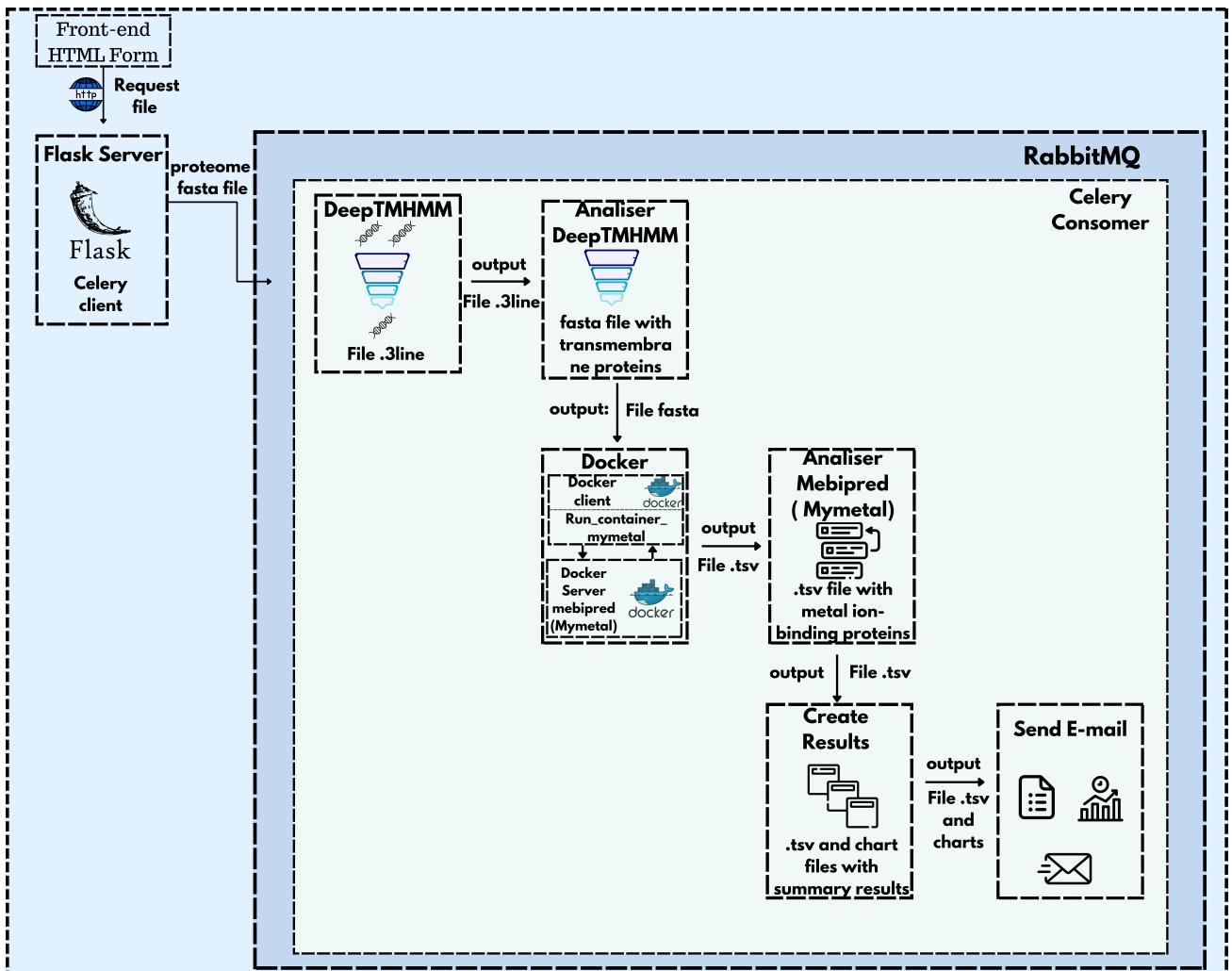


Fonte: Produzido pelo autor (2024).

Um aspecto notável do sistema é sua abordagem simplificada em relação ao registro de usuários. Ao contrário de algumas plataformas que exigem um cadastro extenso, nosso sistema dispensa a necessidade de registro prévio. Os usuários são solicitados apenas a fornecer seu nome completo e um endereço de e-mail válido para receber os resultados da análise. Essa simplicidade visa facilitar o acesso à plataforma e eliminar obstáculos desnecessários para os usuários.

Após o envio da requisição recebida do *Front-end* o pipeline implementado inicia o *workflow*, descrito na seção 5.1.2. Todo o fluxo da arquitetura é ilustrado na Figura 16, onde podemos observar detalhadamente todo o processamento realizado durante a execução. A imagem oferece uma visão abrangente do fluxo de trabalho, facilitando a compreensão do processo em sua totalidade.

Figura 16 – Arquitetura do software



Fonte: Produzido pelo autor (2024).

5.1.1.1 Back-end

O back-end da aplicação “KingFungi” foi desenvolvido utilizando um conjunto de bibliotecas *Python* (versão 3.10.12).

A definição do uso do Python foi feita devido a sua legibilidade e sua sintaxe clara favorecem o entendimento na criação de scripts e programas para automatizar tarefas desafiadoras, o que proporcionam robustez e eficiência na execução das tarefas necessárias para a análise de dados genômicos.

Para o desenvolvimento da aplicação web, optamos por utilizar framework Flask um microframework leve e flexível que facilita a criação de aplicações web e APIs, devido a natureza simples da interface criada no Front-end, criamos apenas uma rota na raiz (“/”), com método “GET” e “POST”, para receber as requisições enviadas pelos usuários e poder renderizar a página. Flask foi escolhido por sua simplicidade e pela facilidade de integração com outras bibliotecas e ferramentas.

Para gerenciar tarefas assíncronas e de longa duração, como o processamento de dados genômicos, foi utilizado o *Celery*, uma biblioteca de distribuição de tarefas que permite a execução eficiente de tarefas em segundo plano. O *Celery*, em conjunto com o *RabbitMQ*, que atua como um broker de mensagens, garante que as tarefas sejam enfileiradas e executadas de maneira ordenada e confiável, o Celery possui um módulo interno chamado *Chain* que possibilita criar essas tarefas sequenciais. *RabbitMQ* facilita a comunicação entre diferentes partes da aplicação, assegurando que as tarefas sejam distribuídas e processadas adequadamente.

A análise dos dados genômicos contidos nos arquivos .fasta foi conduzida utilizando a biblioteca *BioPython*. Esta ferramenta foi fundamental para verificar a validade dos arquivos .fasta fornecidos pelos usuários. Além disso, na etapa 2, conforme descrito na seção 5.1.2, utilizamos o módulo *SeqIO* do *BioPython* para gerar um novo arquivo .fasta. Esse processo destacou a versatilidade da biblioteca no manuseio e na manipulação de dados genômicos.

Para assegurar a continuidade operacional de nossa aplicação, optamos por containerizar a ferramenta *MeBipred* (também conhecida como MyMetal), um componente crucial de nosso pipeline de processamento. Desenvolvemos uma imagem *Docker* específica para o *MeBipred*, fundamentada na versão mymetal 1.0.9, com o propósito de mitigar possíveis contratemplos na execução do referido módulo. Esta imagem encontra-se prontamente disponível no *DockerHub* (<https://hub.docker.com/rey4ssis/mymetal>), acessível à comunidade científica interessada.

5.1.1.2 Front-end

O front-end da aplicação web *KingFungi* foi desenvolvido utilizando tecnologias modernas de desenvolvimento web para proporcionar uma interface funcional e intuitiva aos usuários. A estrutura do documento HTML foi organizada de maneira semântica para garantir a acessibilidade e a manutenção do código, atuando como a principal interface de interação com o usuário. Para garantir um design responsivo e consistente, a biblioteca CSS *Bootstrap* (versão 4.5.2) foi integrada ao projeto. O *Bootstrap* facilita a criação de layouts flexíveis e responsivos através de seu sistema de grid, componentes pré-estilizados e utilitários CSS, com links para os arquivos CSS incluídos diretamente via CDN.

O formulário de submissão de dados genômicos foi projetado para permitir a escolha entre o envio de sequências via texto ou arquivo, utilizando campos de entrada adaptáveis e validados. Este formulário foi enriquecido com *JavaScript* personalizado e *jQuery* para fornecer uma experiência de usuário dinâmica. Scripts *JavaScript* foram usados para alternar entre os campos de texto e arquivo, exibir alertas de sucesso e enviar o formulário via AJAX, evitando o recarregamento da página e melhorando a interatividade.

A inclusão de arquivos CSS personalizados e JavaScript adicional permitiu a estilização específica e o comportamento desejado para a aplicação. A combinação de *HTML5*, *Bootstrap*, *jQuery* e *JavaScript* personalizados resultou em um front-end robusto e eficiente, capaz de oferecer uma interface de usuário intuitiva e responsiva para o envio e análise de sequências genômicas. Essa abordagem garantiu não apenas a funcionalidade desejada, mas também uma experiência de usuário satisfatória, com elementos visuais claros e interativos, facilitando a navegação e o uso da plataforma por pesquisadores e profissionais da área de bioinformática, tornando a análise de genomas acessível mesmo para aqueles com pouca experiência técnica.

5.1.2 Workflow de execução da aplicação

O KingFungi é organizado em duas pastas principais: INPUT e OUTPUT. A pasta INPUT destina-se ao armazenamento dos arquivos a serem analisados, enquanto a pasta OUTPUT recebe as saídas geradas pelo sistema. Cada solicitação feita pelo usuário resulta na criação de uma pasta específica, identificada pelo e-mail do usuário, tanto na pasta INPUT quanto na pasta OUTPUT, onde os arquivos relevantes são manipulados.

O pipeline possui um *workflow* dividido em 6 etapas interligadas para realizar a análise do genoma, e para isso, empregamos o módulo Chain do Celery. Essa estrutura em cadeia permite a execução sequencial e ordenada das etapas, garantindo uma análise eficiente e organizada do genoma. Esse módulo implementa o fluxo de uma pipeline, onde a saída da etapa 1 será a entrada da etapa seguinte, até que o processamento da requisição seja concluído. Assim, em uma situação onde uma das etapas tem apenas uma entrada, não é necessário inseri-la manualmente, pois o Chain automaticamente cuida dessa operação. É importante salientar que, após o início do processamento da pipeline, os dois softwares principais criam, dentro da pasta OUTPUT/EMAIL, duas novas pastas. Uma chamada “Deep” e outra chamada “Mymetal”, onde serão salvos os arquivos das etapas.

ETAPA 1: DeepTMHMM

Nesta etapa, o DeepTMHMM é executado. Na nossa pipeline, ele é utilizado para encontrar em um proteoma todas as proteínas transmembranas, após receber um arquivo .fasta, como o exemplo demonstrado na Figura 17.

O arquivo .fasta é um formato amplamente utilizado em bioinformática para armazenar sequências biológicas, como DNA, RNA ou proteínas. Cada entrada em um arquivo .fasta começa com uma linha de descrição, que é precedida por um símbolo de maior (">") e geralmente contém informações sobre a sequência, como identificador e comentários. Como demonstrado na Figura 17.

Figura 17 – Estrutura de um arquivo .fasta

```

1 >XP_006453741.1 hypothetical protein AGABI20RAFT_39845, partial [Agaricus bisporus var. bisporus H97]
2 VFSKTEKWTIVVIIIAFAGLSPLTANIFYFPAPLTLISIAFKSTELINTVTTMMYILGQIAPMINGTLSNDVGRPVTAAAC
3 LLVLISLCVGLALVPTNAYILLMIIRCLOSAGSASTIAIGAGVSDISTPEERAGFFGFTLGPMPVGPAIGPVIGALAQ
4 GLGWSRFWFLCIAAACLCLMIIISIMPETLTAVERKKDPLSRILYTPLLPIVGRG0ANOPNNIIPPKKLQLNPLKLFTKPD
5 IVLLLAITTAICAVFYGIATSTLAVRYKKDPLSRILYTPLLPIVGRG0ANOPNNIIPPKKLQLNPLKLFTKPD
6 QVDMKDLAKNPPDFPLEHARLVLLPPMIIILAACTAGYGWALEKRVNTACPLILOIIMGYCIGVMNASSTIMIDLVPGQQ
7 SATACCNLVRCLLSAALVSVIQLIIFGIGVGWTYILLTGLTLLSLPVTYLELKIGPRIRKR
8 >XP_006453742.1 hypothetical protein AGABI20RAFT_40608, partial [Agaricus bisporus var. bisporus H97]
9 MKSIFIGLSFISAATAQIVNGLSSVPVITAASASAQSAAATPAPSSAPTPTPQPSGVAQYTPPAQEDFYSYMPYSS
10 MTSGGYSOLQCGYYQKQGDGSCVPLSWQQPONOCYATLFLTYCRGSNYGHCDNSYGSQAATVTVNYVTQTQTVATT
11 VTMEHATATKVETDTMTATKIYTSVEVVPTTRWISTEVIDRTKEIEHTLTATEKTOTNIYTRTATOTDTATOTVKTOTDT
12 ATVTRTDIRTVVQOPTTYIKTYDVTKVIDNTQTWTETATSVTDVKTYLQATATKQTMTDFMTVTAISTMVEPTTYVKTW
13 QTVKTDNTETVLNLNTLSTVTDNQTRLQTVTLLSTATATTTATEIOAAAENTGLSNCLGMCKSSWPTPGYSSYAQPT
14 ATAADDSSSSYGGSGGGSGSYGS6
15 >XP_006453743.1 hypothetical protein AGABI20RAFT_41003, partial [Agaricus bisporus var. bisporus H97]
16 MSFVKLSIFGTSFEVTRYDLDQPVGMGAFLVCSAKDQLTGAVSAIKKIMKPFPSTPVLSKRTYRELKKLKHIOHNEIIS
17 AVDFVISPLDEDVTELLSRPLEQFQYFLYIQLRLKYVHSAGVWHRDLKPSNIVNQFQIICLICDFG
18 LARIQDPQMTGYVSTRYYRAPEIIMLWKOYDVAVDIWSAGCIFAEMLLEGKPLFPKGDKHVNQFQSIITELLGTPDDVIEITI
19 CSSENTLRFVQSLPKRDRQPFSEKLRSTDPAEALLLECMVLFDPRKRIDAADSLAHSYVAPYHDPTDEPVAEKFDFWSFND
20 ADLPVDTWKVMMYSEILDHFHQVGD

```

Fonte: Produzido pelo autor (2024).

O DeepTMHMM gera 3 arquivos de saída: um arquivo de anotação .gff3, um arquivo .3line e um arquivo .md, contendo todas as informações da análise feita no software. Optamos por utilizar o arquivo .3line, onde o software imprime o resultado do processamento, indicando o código de acesso da proteína, um termo de identificação do tipo e a sequência da proteína. Abaixo da sequência, para cada aminoácido, uma letra representa sua localização, identificando a parte que corresponde à membrana (M), à parte interna (I) e à parte externa (O), que significam “Membrane”, “Inside” e “Outside”, respectivamente.

Existem 3 possíveis termos que indicarão se a proteína em questão é ou não transmembrana:

Termo 1: TM - Esse termo indica que se trata de uma proteína transmembrana helicoidal sem peptídeo sinal.

Termo 2: SP+TM - Esse termo indica que se trata de uma proteína transmembrana com peptídeo sinal.

Termo 3: BETA - Indicação de que é uma proteína transmembrana de barril beta.

Como podemos ver na Figura 18:

Figura 18 – Exemplo do arquivo de saída do DeepTMHMM(.3line)

Fonte: Produzido pelo autor (2024).

ETAPA 2: Analisador do resultado do DeepTMHMM

Para identificar as proteínas-alvo, analisamos o arquivo .3line, observando se a proteína é transmembrana através do termo impresso no arquivo. Se um dos 3 termos for encontrado (TM, SP+TM ou BETA), selecionamos o código da proteína e a sequência e geramos um novo arquivo .fasta, que é salvo dentro da pasta "Deep" com o nome "deep_out.fasta", demonstrado na Figura 19. Esse arquivo também é retornado no código construído para essa etapa.

Figura 19 – Exemplo do arquivo de saída do analisador DeepTMHMM (deep_out.fasta)

1 >XP_006453741.1
2 VFSKTEKWIVVIAFAGLFSPLTANIIYFPAPITLSIAFKSTELINLTVMYMILOQIAPMIWGTLSNVGRRPVTACLLV
3 LSLSCVGLALVPTNAYWLLMILRCLQSAGSASTIAIGAGVVSIDTPEERAGFFGFTLGPVMVGAIGPVIGGALAQQLGLWRSI
4 FWFLCIAAAALCLVMIISIMPETAVRYKKDPLSRILYTPPLPIVGRGQANQPNNIPPKKLQNPLKLFKPDIVLLAITA
5 ICAVFGIIASISTLFNDTYDLYNVTTLGLCFLAIGGGMAIGSSINGRIMDKWFEKEKQFAKDMAAGKQVDMKDLAKNPDF
6 PLEHARLVLLPPMIIILACTAGYGWALEKRVNIA CPLILQIIMGYMCIGVMNASSTIMIDL VPGQGSAITACNNLVRCLLSA
7 ALVSVIQOLIFDGGIGVGWTYILLTGTLTLLSPVTYLELKIGPRIRK
8 >XP_006453764.1
9 MTPLAPNRKLSAEQVEQSARDPTAALYDAVARSPPRPVEPLPTHIDISRTASPIPMAARFSDSHLPPLDDSVVTRWRVLLYL
10 ISTTALWVFGGIFFMLLGIGLTVQVIIFPRIHTCPSSATCENSYDPNNGSNLOLQSFMSYWLKAGLMIASVGILKLAAYQAWF
11 ILMHEGNTVKDLDLNLGAIRGSVTDAFFLFRKHNRLLSIVFALLGVDTAISLIIGLSINKQSGTKVDFYNATSRFPDSS
12 LSHLNSDQLKATQKSIWALDGDKSHGGALRGSLVVPGDRSIQATNALPAGPKISGRFECGYSNTYFDPESSPLHQWYINV
13 DRDQYIANAKMSLHVAMHVDTAVTRYLWVNSTTGLIPNATATEDGGMHIAFCTHWEMVPEEPKKAGYDLYNAAFTSGCS
14 DEAGSETCVADSVNNAILNWGGVGTAFWHISCRGGVLGPVPSRNDAERYCSLTQELWKETTIAMLDGITQTAPTSIPSSQKL
15 QAGVEELNRQRWLNNAVIPAAIFVLYLVGLVYTSLRSQGNPAPKKLNDEVRAAQTDIHDLILMGQLKTPVRYHSQIGFV
16 DSHNHSGT

Fonte: Produzido pelo autor (2024).

ETAPA 3: Mebiprev(Mymetal)

O código executado dentro do container receberá como parâmetros um arquivo .fasta, uma pasta de entrada e uma pasta de saída.

Utilizamos a biblioteca Docker SDK para criar um container através da imagem disponibilizada, passando as seguintes informações:

- Arquivo de entrada: o arquivo de saída da etapa anterior (deep_out.fasta).
- Input: OUTPUT/EMAIL/Deep
- Output: OUTPUT/EMAIL/Mymetal

Essa etapa resultará em um arquivo .TSV que será salvo na pasta de saída especificada no parâmetro. Esse arquivo conterá informações sobre quais proteínas se ligam aos elementos descritos na subseção 4.1.1 Mebipred (Mymetal) do Capítulo 4.

ETAPA 4: Analisador do resultado do Mebiprev(Mymetal)

Nesta etapa, analisaremos o arquivo .TSV da etapa 3, onde é necessário aplicar um filtro entre as proteínas encontradas. Esse filtro tem como objetivo estabelecer um critério mais rigoroso, considerando que se trata de uma predição. Portanto, é crucial garantir um alto grau de confirmação. Para isso, determinamos um score ≥ 0.8 . Se a proteína atingir o score de confiança para cada elemento, é adicionado o valor inteiro “1” para confirmação e “0” para a negação. Ao final, obteremos um novo arquivo .TSV com 11 colunas.

- Coluna 1: ID da proteína;
- Coluna 2: Proteína que se liga a Cálcio;
- Coluna 3: Proteína que se liga a Cobalto;
- Coluna 4: Proteína que se liga a Cobre
- Coluna 5: Proteína que se liga a Ferro
- Coluna 6: Proteína que se liga a Potássio
- Coluna 7: Proteína que se liga a Magnésio
- Coluna 8: Proteína que se liga a Manganês
- Coluna 9: Proteína que se liga a Sódio
- Coluna 10: Proteína que se liga a Níquel
- Coluna 11: Proteína que se liga a Zinco

O arquivo é também salvo na pasta OUTPUT/EMAIL/Mymetal e retornado no código, tornando-se entrada para a próxima etapa. Um exemplo do arquivo pode ser visualizado na Figura 20.

Figura 20 – Exemplo do arquivo de saída do analisador Mymetal

| | ID protein | Cálcio(Ca) | Cobalto(Co) | Cobre(Cu) | Ferro(Fe) | Pótassio(K) | Magnésio(Mg) | Manganês(Mn) | Sódio(Na) | Níquel(Ni) | Zinco(Zn) |
|---|------------|------------|-------------|-----------|-----------|-------------|--------------|--------------|-----------|------------|-----------|
| 1 | KZV7782.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | KZV77804.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | KZV77812.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | KZV77822.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Fonte: Autor.

ETAPA 5: Criador dos resultados

Nesta etapa, criamos os resultados ao receber o arquivo .TSV retornado na etapa anterior e construímos o gráfico a partir dos dados contidos nele. Para isso, utilizamos a biblioteca Pandas do Python para ler e formatar os dados do arquivo .TSV. Além disso, empregamos a biblioteca Matplotlib.pyplot para criar um gráfico de barras verticais. Optamos por este tipo de gráfico devido à sua capacidade de representação clara e concisa, facilitando a visualização dos resultados da análise.

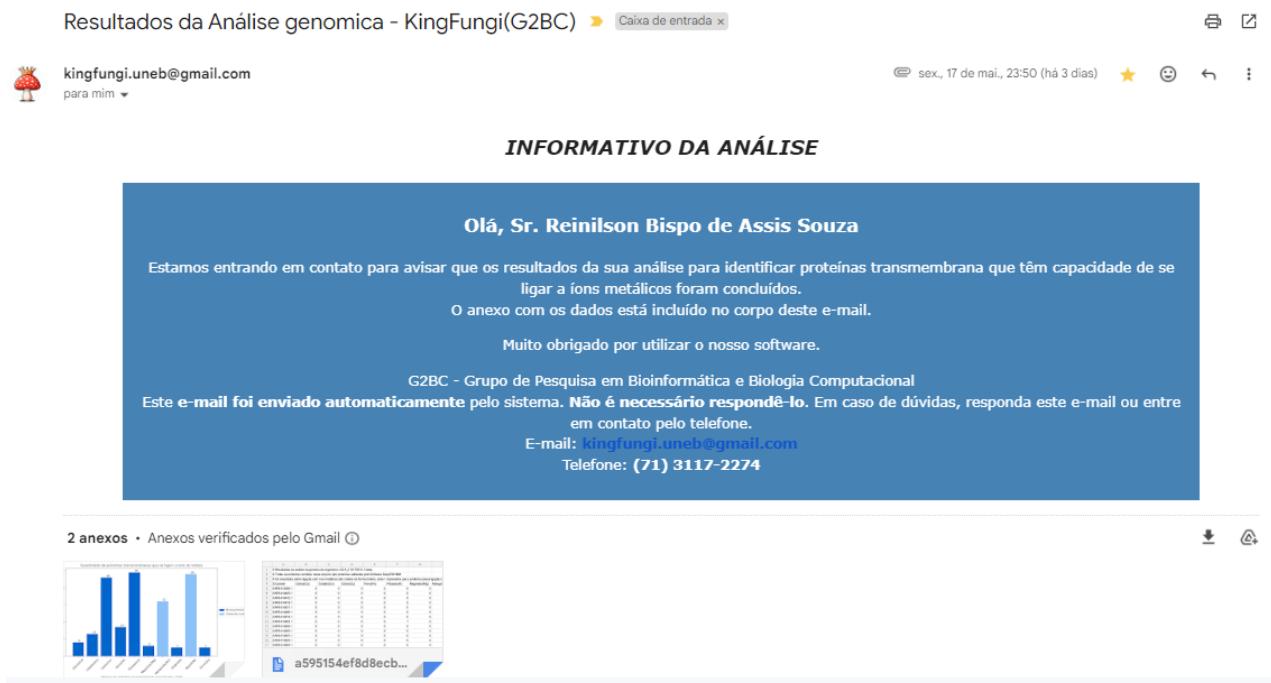
Adicionamos três linhas de comentários no arquivo .TSV, além do conteúdo já existente. São elas:

- Identificação do organismo analisado com o Refseq.
- Explicação de que todas as proteínas contidas neste arquivo são transmembranas.
- Explicação do formato do arquivo: no arquivo, o valor “1” indica que a proteína possui ligação com aquele elemento, enquanto o valor “0” indica que não há ligação. Os dados são desse arquivo são utilizados para gerar o gráfico.

ETAPA 6: Enviar e-mail com os resultados

Por fim, na etapa 6, a etapa 6 recebe o gráfico e o arquivo .TSV criado na etapa anterior e os enviaremos para o e-mail fornecido pelo usuário durante a criação da requisição. Essa mensagem contém dois anexos: o gráfico gerado com os resultados e o arquivo .TSV correspondente. Na Figura 21, temos um exemplo do e-mail contendo o resultado da análise.

Figura 21 – E-mail de resultados



Fonte: Produzido pelo autor (2024).

5.2 Avaliação de eficiência (tempo de execução) e eficácia (funcionalidade)

Nesta seção serão apresentados as informações sobre as duas avaliações propostas na metodologia DSR, qualitativa em termos de funcionalidade da aplicação web e quantitativa em termos de tempo de execução da aplicação web.

5.2.1 Ambiente de avaliação e genomas utilizados

Para validar o aplicação desenvolvida, empregamos dados genômicos obtidos do GenBank. O GenBank é uma das mais abrangentes bases de dados genômicos, oferecendo uma vasta coleção de sequências genéticas de diversos organismos. Para assegurar uma análise robusta, selecionamos genomas de cogumelos comestíveis de diferentes ordens pertencentes ao filo Basidiomycota garantindo uma amostragem diversificada e representativa, essencial para avaliar a capacidade da aplicação de lidar com diferentes perfis genômicos, assim como (Scheid *et al.*, 2020), que desenvolveu uma pesquisa recentemente que demonstrou taxas relevantes de acumulação de ferro em cogumelos comestíveis do filo Basidiomycota.

Os genomas utilizados foram de três espécies distintas com informações disponíveis na Tabela 4:

Tabela 4 – Organismos selecionados para a avaliação do KingFungi (KF)

| Ordem | Espécie selecionada | Quant. de proteínas | Quant. de pares de bases (Mb) | Número de acesso GenBank |
|-----------------------|--------------------------|---------------------|-------------------------------|--------------------------|
| <i>Agaricales</i> | <i>Agaricus bisporus</i> | 10.448 | 30.4 Mb | GCA_000300575.1 |
| <i>Auriculariales</i> | <i>Exidia glandulosa</i> | 26.690 | 78.2 Mb | GCA_001632375.1 |
| <i>Boletales</i> | <i>Boletus edulis</i> | 18.718 | 66.5 Mb | GCA_015179015.1 |

Fonte: Produzido pelo autor (2024).

Os experimentos foram conduzidos em um Desktop equipado com as seguintes configurações:

- Sistema operacional: Ubuntu versão 22.04.4 LTS-64 bits
- Memória RAM: 32GB de RAM
- Processador: AMD Ryzen 9-7900 4,7 GHz (com 12 cores e 24 threads)
- Armazenamento: 5TB, distribuído entre 2 HDDs de 2TB e um SSD de 1TB

5.2.2 Avaliação 1 (Funcionalidades)

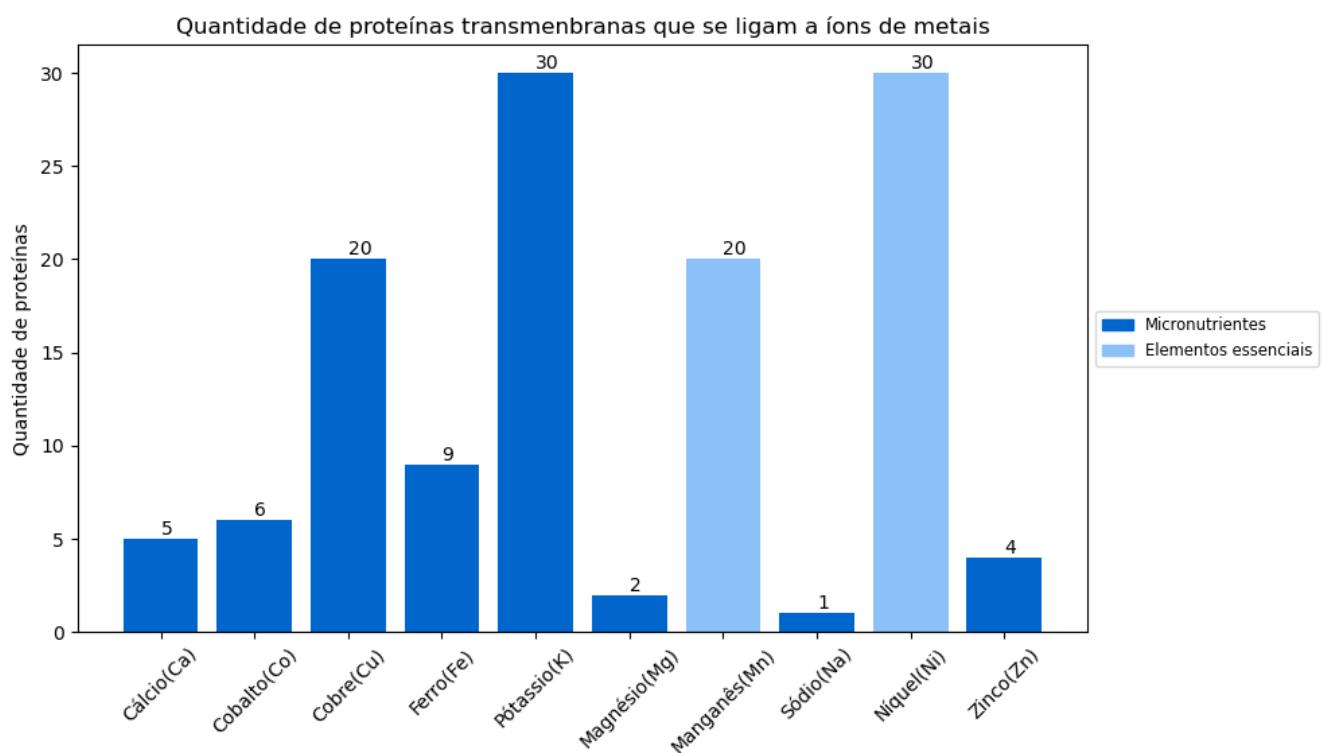
Os dados serão visualizados por meio de gráficos de barras verticais, que demonstram o número de proteínas transmembrana associadas a cada elemento demonstrados nos Gráficos 22, 24, 26. Além disso, as 10 primeiras linhas do arquivo .TSV estão apresentadas nas Figuras 23 25 27.

Os elementos estudados abrangem um total de 10 metais, dos quais 8 são micronutrientes e 2 são elementos essenciais. Os micronutrientes analisados incluem: Ca, Co, Cu, Fe, K, Mg, Na, Zn. Além desses, foram investigados dois elementos essenciais: Mg e Ni.

1. Agaricales *Agaricus bisporus*

Na análise realizada com o genoma do *Agaricus bisporus*, foram encontradas 1536 proteínas transmembranas. Selecionando apenas as que possuem capacidade de se ligar aos íons metálicos, foram previstas as seguintes quantidades de proteínas: Ca (5), Co (6), Cu (20), Fe (9), K (30), Mg (2), Na (1) e Zn (4), no que diz respeito aos micronutrientes. Em relação aos elementos essenciais, encontramos proteínas que se ligam ao Mn (20) e ao Ni (30).

Figura 22 – Gráfico do *Agaricales Agaricus bisporus*



Fonte: Produzido pelo autor (2024).

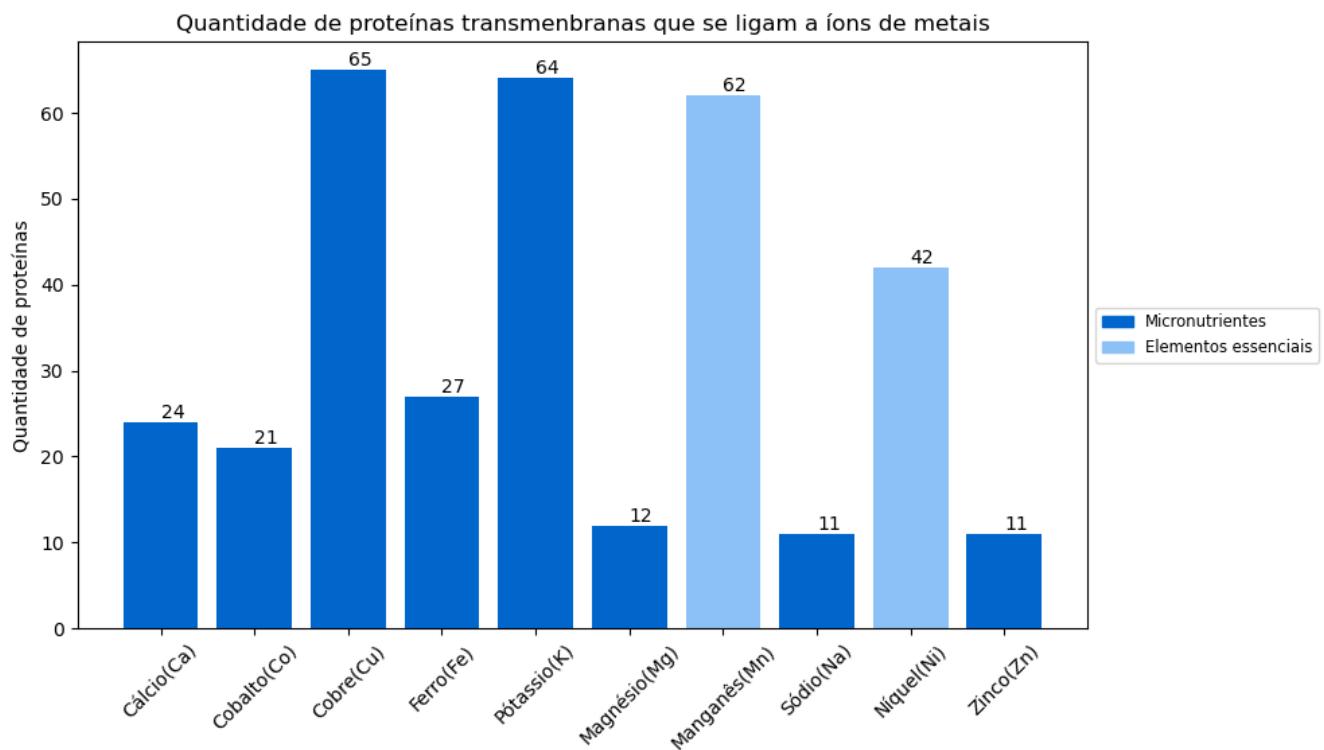
Figura 23 – Resultados da análise no arquivo TSV do *Agaricales Agaricus bisporus*

| ID protein | Cálcio(Ca) | Cobalto(Co) | Cobre(Cu) | Ferro(Fe) | Pótassio(K) | Magnésio(Mg) | Manganês(Mn) | Sódio(Na) | Niquel(Ni) | Zinco(Zn) |
|-------------------|------------|-------------|-----------|-----------|-------------|--------------|--------------|-----------|------------|-----------|
| 1 XP_006453888.1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 2 XP_006453957.1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 3 XP_006454022.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 4 XP_006454114.1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 5 XP_006454288.1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 6 XP_006454289.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 XP_006454612.1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 XP_006454628.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 9 XP_006454677.1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 10 XP_006454703.1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 11 XP_006454729.1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |

Fonte: Produzido pelo autor (2024).

2. Auriculariales *Exidia glandulosa*

Na análise conduzida com o genoma do *Auriculariales Exidia glandulosa*, foram identificadas 3525 proteínas transmembranas. Dentre estas, observou-se a presença de proteínas com afinidade para se ligar aos seguintes íons metálicos: Ca (24), Co (21), Cu (65), Fe (27), K (64), Mg (12), Na (11) e Zn (11), abordando os micronutrientes. Em relação aos elementos essenciais, foram identificadas proteínas que interagem com o Mn (62) e o Ni (42).

Figura 24 – Gráfico do *Auriculariales Exidia glandulosa*

Fonte: Produzido pelo autor (2024).

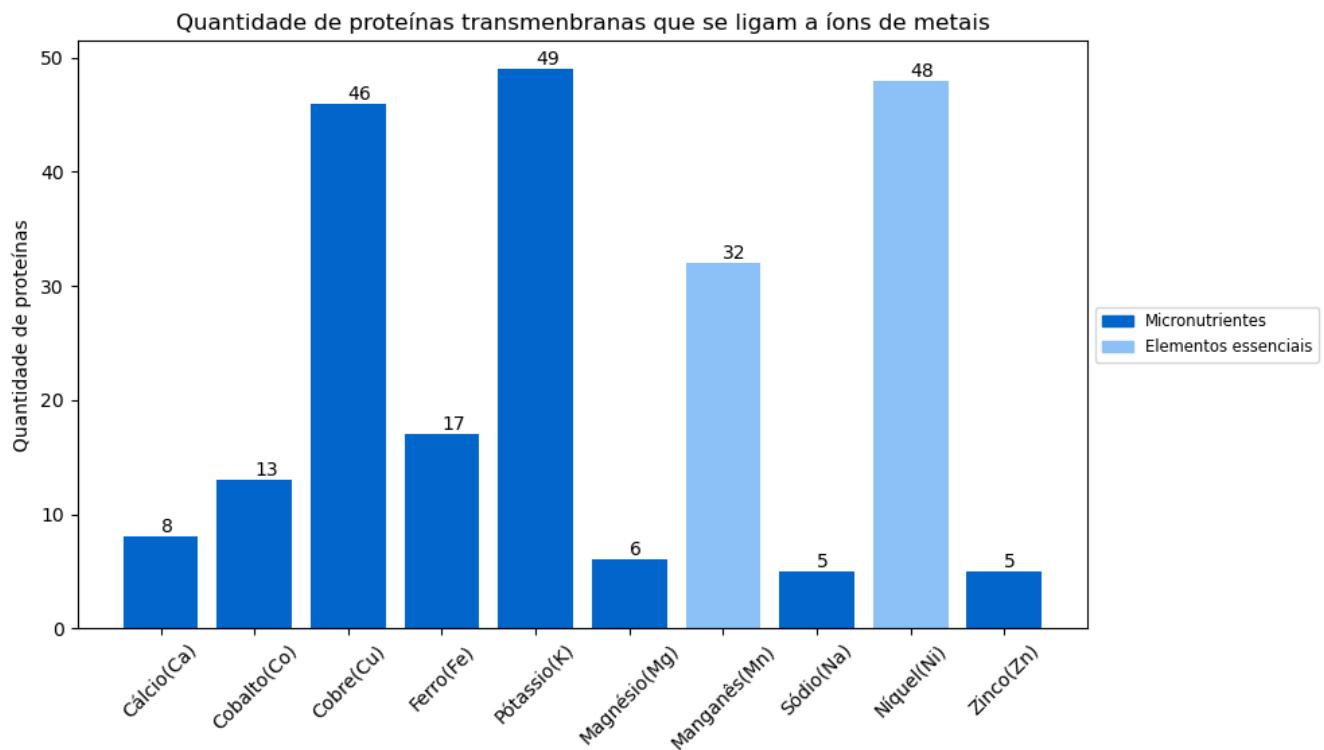
Figura 25 – Resultados da análise no arquivo TSV do *Auriculariales Exidia glandulosa*

| 1 | ID protein | Cálcio(Ca) | Cobalto(Co) | Cobre(Cu) | Ferro(Fe) | Pótassio(K) | Magnésio(Mg) | Manganês(Mn) | Sódio(Na) | Níquel(Ni) | Zinco(Zn) |
|----|------------|------------|-------------|-----------|-----------|-------------|--------------|--------------|-----------|------------|-----------|
| 2 | KZV78013.1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 3 | KZV78042.1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 4 | KZV78131.1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 5 | KZV78925.1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 6 | KZV79541.1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | KZV79599.1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | KZV79603.1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 9 | KZV79636.1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | KZV79785.1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 11 | KZV79952.1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

Fonte: Produzido pelo autor (2024).

3. Boletales *Boletus edulis*

Explorando o genoma do *Boletales Boletus edulis*, foram identificadas 2308 proteínas transmembranas. Dentro desse conjunto, foram observadas proteínas com habilidade de se ligar a íons metálicos, incluindo Ca (8), Co (13), Cu (46), Fe (17), K (49), Mg (6), Na (5) e Zn (5), cobrindo uma variedade de micronutrientes. Quanto aos elementos essenciais, encontraram-se proteínas envolvidas na ligação com Mn (32) e Ni (48).

Figura 26 – Gráfico do *Boletales Boletus edulis*

Fonte: Produzido pelo autor (2024).

Figura 27 – Resultados da análise no arquivo TSV do *Boletales Boletus edulis*

| 1 | ID protein | Cálcio(Ca) | Cobalto(Co) | Cobre(Cu) | Ferro(Fe) | Pótassio(K) | Magnésio(Mg) | Manganês(Mn) | Sódio(Na) | Níquel(Ni) | Zinco(Zn) |
|----|--------------|------------|-------------|-----------|-----------|-------------|--------------|--------------|-----------|------------|-----------|
| 2 | KAF8414483.1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 3 | KAF8414704.1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | KAF8415711.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 5 | KAF8416475.1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 6 | KAF8418923.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 7 | KAF8423174.1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | KAF8423727.1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 9 | KAF8424416.1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | KAF8425747.1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 11 | KAF8428050.1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Fonte: Produzido pelo autor(2024).

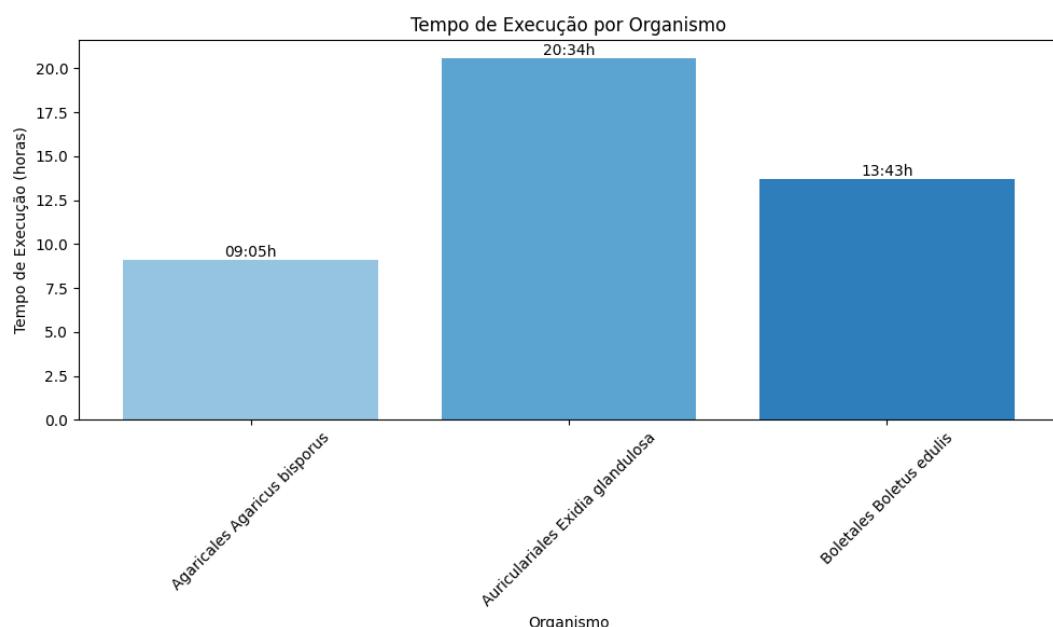
Após avaliação qualitativa, confirmamos que a aplicação *KingFungi* foi eficaz na identificação das proteínas transmembrana que se ligam a micronutrientes e outros elementos essenciais. Esses resultados validam satisfatoriamente o objetivo principal da nossa pesquisa, evidenciando claramente a capacidade dessa ferramenta em contribuir significativamente para o avanço do conhecimento em biologia molecular.

5.2.3 Avaliação 2 (Tempo)

É importante notar que o tempo de execução pode variar conforme o ambiente utilizado para a realização dos testes.

O tempo necessário para a execução apresentada na seção anterior foi registrado em horas e está disponível para visualização no gráfico 28. As análises obtiveram um tempo médio de 14 horas e 27 minutos.

Figura 28 – Tempo de execução da análise dos 3 genomas



Fonte: Produzido pelo autor (2024).

O tempo de execução da aplicação atualmente é significativo, podendo afetar a experiência do usuário. Para aprimorar a performance, foram implementados ajustes no código.

5.3 Remodelagem do pipeline

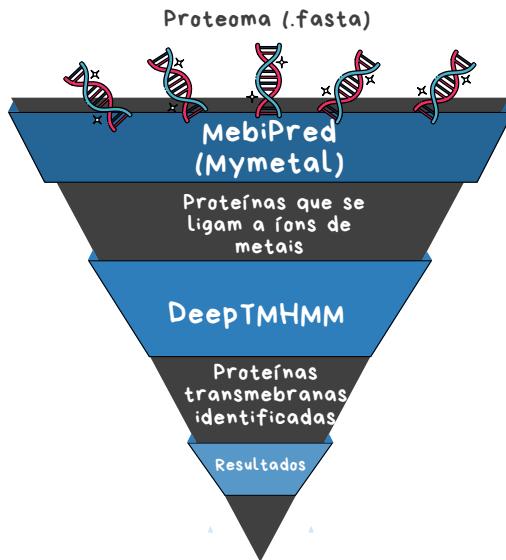
A metodologia de desenvolvimento definida e utilizada neste projeto, descrita na seção 4.1, o DSR, que visa avaliar continuamente o artefato desenvolvido, nos possibilitou ‘enxergar’ uma limitação em termos de tempo de execução, pois a média de tempo descrita na seção 5.2.3 demonstrou um tempo relativamente alto. Para resolver isso, reorganizamos a ordem do pipeline. Percebemos, durante a validação, que, enquanto o DeepTMHMM realizou a análise do primeiro genoma (*Agaricus bisporus*, com 1448 sequências de proteínas) em um tempo de 09:05h, o Mebipred, de forma isolada (fora do pipeline), realizou a análise buscando proteínas que se ligam a íons de metais e resultou em um tempo de aproximadamente 21 segundos, como demonstrado na Figura 29. Logo, concluímos que deveríamos reorganizar o pipeline para que o Mebipred fosse o primeiro software. Com base nos resultados dele no arquivo .TSV, utilizamos o Biopython para gerar um arquivo .fasta apenas com as proteínas que se ligam a metais, para que, no DeepTMHMM, sejam identificadas quais são transmembranas. A Figura 30 demonstra a remodalagem do pipeline.

Figura 29 – Tempo de execução do MebiPred de forma isolada

```
● rey@rey:~/Documentos/SALA DE TESTES$ ./bin/python3 "/home/rey/Documentos/SALA DE TESTES/Teste tempo/appT.py"
2024-06-02 13:19:04.636655: E external/local_xla/xla/stream_executor/cuda/cuda_dnn.cc:9261] Unable to register
when one has already been registered
2024-06-02 13:19:04.636705: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:607] Unable to register
when one has already been registered
2024-06-02 13:19:04.637903: E external/local_xla/xla/stream_executor/cuda/cuda_blas.cc:1515] Unable to register
LAS when one has already been registered
2024-06-02 13:19:04.644268: I tensorflow/core/platform/cpu_feature_guard.cc:182] This TensorFlow binary is opt
l operations.
To enable the following instructions: AVX2 FMA, in other operations, rebuild TensorFlow with the appropriate c
2024-06-02 13:19:05.267239: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not
Progress bar disabled for multiprocesing large sample.
2024-06-02 13:19:19.902042: E external/local_xla/xla/stream_executor/cuda/cuda_driver.cc:274] failed call to c
327/327 [=====] - 0s 1ms/step
Tempo total de execução: 20.999882698059082
```

Fonte: Produzido pelo autor (2024).

Figura 30 – Representação gráfica das etapas do pipeline pós remodelagem



Fonte: Produzido pelo autor (2024).

Após a remodelagem, foram feitas novas avaliações para confirmar o funcionamento e, assim, garantir a eficácia. Essas verificações foram cruciais para integrar as mudanças de maneira harmônica e manter o sistema operando de maneira eficiente.

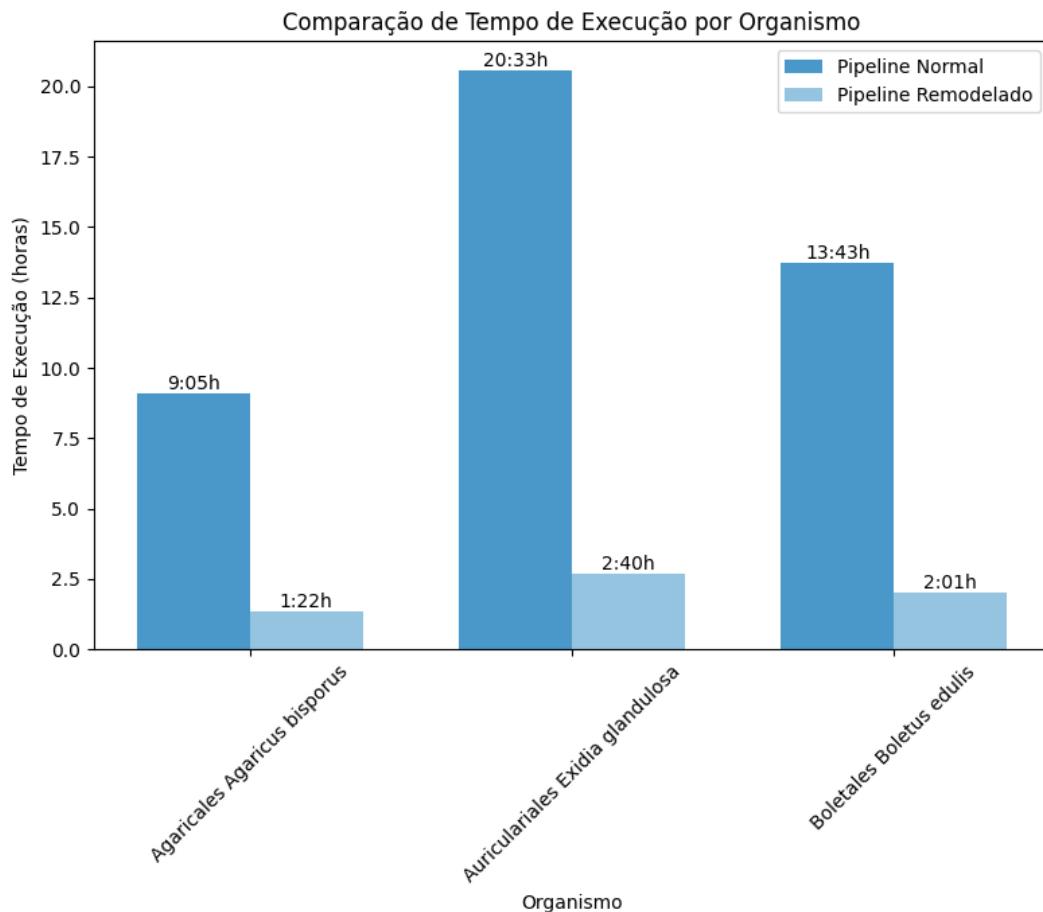
5.3.1 Avaliação 1 (Funcionalidades) pós remodelagem

Após a remodelagem e organização do pipeline realizamos novamente a etapa de avaliação com os mesmos genomas definidos anteriormente na seção 5.2.1. A segunda avaliação não mostrou alterações na quantidade das proteínas que se ligam a íons metálicos, o que confirmar a capacidade da aplicação de realizar tais análises.

5.3.2 Avaliação 2 (Tempo) pós remodelagem

A remodelagem da pipeline serviu para melhorar significativamente o tempo das análises, otimizando o processamento de dados e permitindo uma execução mais eficiente das tarefas. A comparação entre os tempos de execução, pode ser visto na Figura 31.

Figura 31 – Tempo de execução da análise dos 3 genomas pós remodelagem



Fonte: Produzido pelo autor(2024).

Como resultado, a eficiência operacional foi aumentada, possibilitando a obtenção de *insights* mais rápidos. O tempo médio das análises diminuiu para aproximadamente 2h (pós remodelagem).

6 TRABALHOS FUTUROS

Para melhorar a experiência do usuário e tornar a navegação mais intuitiva, é essencial implementar um recurso de autocompletar no campo de inserção do nome do organismo no sistema. Isso permitirá que os usuários localizem rapidamente o organismo desejado, economizando tempo e esforço durante a utilização da aplicação. Além disso, devido à complexidade e ao tempo de processamento exigidos pela inteligência artificial para análise de dados, é fundamental otimizar esse processo. Uma estratégia eficaz seria a criação de um banco de dados centralizado contendo informações dos genomas já analisados. Essa medida não só reduzirá o tempo necessário para obter resultados, mas também evitaria redundâncias ao repetir análises já realizadas.

Em uma pesquisa recente, o grupo *Genome Canada* desenvolveu ferramentas alinhadas com o objetivo deste trabalho. A ferramenta *ToiT-Proteome*, por exemplo, é capaz de identificar proteínas transportadoras e suas propriedades em um proteoma. Além disso, a ferramenta *ToiT-BERT_ICAT* prevê e classifica proteínas transportadoras de íons inorgânicos com alta precisão e exatidão. Para trabalhos futuros, seria interessante comparar os resultados obtidos pelo *KingFungi* com os das ferramentas mencionadas.

A aplicação não possui a capacidade de identificar o micronutriente selênio, que desempenha um papel crucial como antioxidante, catalisador hormonal e fortalecedor do sistema imunológico. Além de contribuir para a motilidade dos espermatozoides e reduzir o risco de aborto espontâneo, sua deficiência está associada a estados de humor adversos. Uma ingestão adequada pode até mesmo reduzir o risco de câncer. Logo, é necessário expandir o sistema, encontrando um novo software para atuar no pipeline que consiga identificar esse micronutriente.

7 CONSIDERAÇÕES FINAIS

Foram definidos 3 OKRs para serem desenvolvidos nesse trabalho: a realização de uma revisão de literatura, a proposição de um método computacional para a identificação dos genes alvo nos genomas de cogumelos e a avaliação da eficiência do método em termos de tempo de execução para genomas de cogumelos de diferentes ordens. Os resultados obtidos ao longo desses testes são altamente congruentes com as expectativas. Foi desenvolvida a aplicação Web, *KingFungi*, que mostrou-se capaz de identificar proteínas transmembrana em um genoma e de discernir quais dessas proteínas têm a capacidade de se ligar a íons de metais, incluindo micronutrientes e elementos essenciais para a saúde humana. Com base na condução da validação da aplicação *KingFungi*, percebe-se que esta etapa não apenas se revela crucial, mas também assume um papel essencial na confirmação da eficiência e robustez do método proposto. Ao utilizar dados do *GenBank*, o pipeline demonstrou sua adaptabilidade diante de uma diversidade de perfis genômicos de diferentes ordens do filo *Basidiomycota*. Esta inovação potencialmente representa um avanço significativo na área da Bioinformática, reduzindo consideravelmente o tempo e esforço necessários para análises complexas.

O *KingFungi* foi validado em duas etapas distintas com 3 genomas de cogumelos comestíveis escolhidos. Na primeira etapa, utilizando o pipeline em uma ordem definida inicialmente, obteve-se um tempo médio de processamento de 14 horas e 27 minutos. Após sua remodelagem, o tempo de processamento foi reduzido para aproximadamente 2 horas. Com a capacidade de identificar até 10 elementos em uma única requisição, sendo eles cálcio, cobalto, cobre, ferro, potássio, magnésio, sódio e zinco que são micronutrientes e manganês, níquel que são elementos essenciais.

A validação adicional revelou que a aplicação não apenas suporta a análise de arquivos locais, mas também permite o processamento com base apenas no nome ou RefSeq do organismo de interesse obtido do *Genbank*. Esse recurso flexível aumenta ainda mais a utilidade e praticidade do *KingFungi*, proporcionando uma abordagem acessível e eficiente para uma variedade de pesquisadores e instituições.

Além disso, o *KingFungi* pode ter um impacto significativo no combate à fome e insegurança alimentar. Novas espécies de cogumelos, identificadas e caracterizadas com a ajuda da ferramenta, podem ser cultivadas em espaços pequenos, sem grandes investimentos, em áreas urbanas. Essas espécies podem ser muito úteis para combater a insegurança alimentar, fornecendo alimentos nutritivos e essenciais para a saúde humana. A aplicação eficiente do *KingFungi* pode, portanto, não apenas avançar a pesquisa científica, mas também contribuir para soluções práticas e sustentáveis na produção de alimentos em

comunidades urbanas.

REFERÊNCIAS

- AHMAD, M. O.; MARKKULA, J.; OIVO, M. Kanban in software development: A systematic literature review. In: **2013 39th Euromicro conference on software engineering and advanced applications**. [S.l.]: IEEE, 2013. p. 9–16. Citado na página 56.
- AL-OBEIDAT, F. *et al.* (CDRGI)-Cancer detection through relevant genes identification. **Neural Comput & Applic**, v. 34, n. 11, p. 8447–8454, jun. 2022. ISSN 1433-3058. Disponível em: <https://doi.org/10.1007/s00521-021-05739-8>. Citado na página 45.
- ALBERTS, B. *et al.* Ion channels and the electrical properties of membranes. In: **Molecular Biology of the Cell. 4th edition**. [S.l.]: Garland Science, 2002. Citado na página 30.
- ALBERTS, B. *et al.* **Biologia molecular da célula**. [S.l.]: Artmed Editora, 2017. 299-365 p. Citado nas páginas 25, 26, 29 e 30.
- ANGUITA-RUIZ, A. *et al.* explainable artificial intelligence (xai) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. **PLoS computational biology**, Public Library of Science San Francisco, CA USA, v. 16, n. 4, p. e1007792, 2020. Citado na página 14.
- APTEKMANN, A. A. *et al.* mebipred: identifying metal-binding potential in protein sequence. **Bioinformatics**, Oxford University Press, v. 38, n. 14, p. 3532–3540, 2022. Citado nas páginas 28, 47, 49 e 50.
- ARIF, L. *et al.* Role of micronutrients (vitamins & minerals). **International Journal of Multidisciplinary Sciences and Arts**, v. 3, n. 01, 2024. Citado na página 23.
- BALTZ, R. H. Genome mining for drug discovery: progress at the front end. **Journal of Industrial Microbiology and Biotechnology**, Oxford University Press, v. 48, n. 9-10, p. kuab044, 2021. Citado na página 13.
- BARBOSA, J. R. *et al.* Polysaccharides of mushroom pleurotus spp.: New extraction techniques, biological activities and development of new technologies. **Carbohydrate polymers**, Elsevier, v. 229, p. 115550, 2020. Citado na página 14.
- BARTLEY, W.; DAVIES, R. E. Active transport of ions by sub-cellular particles. **Biochemical Journal**, Portland Press Ltd, v. 57, n. 1, p. 37, 1954. Citado na página 29.
- BAUMAN, K. D. *et al.* Genome mining methods to discover bioactive natural products. **Natural Product Reports**, Royal Society of Chemistry, v. 38, n. 11, p. 2100–2129, 2021. Citado na página 32.
- BEELMAN, R. B.; KALARAS, M. D.; JR, J. P. R. Micronutrients and bioactive compounds in mushrooms: a recipe for healthy aging? **Nutrition Today**, LWW, v. 54, n. 1, p. 16–22, 2019. Citado na página 13.
- BELL, V. *et al.* Mushrooms as future generation healthy foods. **Frontiers in Nutrition**, Frontiers Media SA, v. 9, p. 1050099, 2022. Citado nas páginas 19 e 20.

- BENSON, D. A. *et al.* Genbank. **Nucleic acids research**, Oxford University Press, v. 41, n. D1, p. D36–D42, 2012. Citado na página 39.
- BERGER, M. M. *et al.* Espen micronutrient guideline. **Clinical Nutrition**, Elsevier, v. 41, n. 6, p. 1357–1424, 2022. Citado na página 24.
- BEZERRA, T. A.; OLINDA, R. A. d.; PEDRAZA, D. F. Insegurança alimentar no brasil segundo diferentes cenários sociodemográficos. **Ciência & Saúde Coletiva**, SciELO Brasil, v. 22, p. 637–651, 2017. Citado na página 17.
- BILLS, G. F.; GLOER, J. B. Biologically active secondary metabolites from the fungi. **Microbiology spectrum**, Am Soc Microbiol, v. 4, n. 6, p. 4–6, 2016. Citado na página 21.
- BISANÇÃO, V. R.; POLONIO, J. C.; GOLIAS, H. C. Cogumelos basidiomycota: Fontes de compostos com atividade anticâncer. **Arquivos do Mudi**, v. 26, n. 2, p. 29–46, 2022. Citado na página 26.
- BITO, T. *et al.* Characterization of vitamin b12 compounds in the fruiting bodies of shiitake mushroom (*lentinula edodes*) and bed logs after fruiting of the mushroom. **Mycoscience**, The Mycological Society of Japan, v. 55, n. 6, p. 462–468, 2014. Citado na página 15.
- BOCHIE, K. *et al.* Detecção de ataques a redes iot usando técnicas de aprendizado de máquina e aprendizado profundo. In: SBC. **Anais do XX Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais**. [S.l.], 2020. p. 257–270. Citado na página 51.
- BOURGEY, M. *et al.* Genpipes: an open-source framework for distributed and scalable genomic analyses. **Gigascience**, Oxford University Press, v. 8, n. 6, p. giz037, 2019. Citado na página 31.
- BRISTOT, G. Biomarcadores periféricos e regulação transcricional nos transtornos psiquiátricos: vias moleculares associadas à psicopatologia no transtorno bipolar, transtorno depressivo maior e esquizofrenia. 2022. Citado na página 33.
- BROCKE, J. vom; HEVNER, A.; MAEDCHE, A. Introduction to design science research. **Design science research. Cases**, Springer, p. 1–13, 2020. Citado na página 54.
- CAO, Y. *et al.* Ensemble deep learning in bioinformatics. **Nature Machine Intelligence**, Nature Publishing Group UK London, v. 2, n. 9, p. 500–508, 2020. Citado na página 37.
- CARDOSO, R. P. Objectives and key results (okr) aplicado a uma empresa industrial: um estudo de caso. 2020. Citado na página 56.
- CHEEK, M. *et al.* **New scientific discoveries: plants and fungi. Plants People Planet 2 (5): 371–388.** 2020. Citado na página 19.
- CHEN, H.; DONG, F.; MINTEER, S. D. The progress and outlook of bioelectrocatalysis for the production of chemicals, fuels and materials. **Nature Catalysis**, Nature Publishing Group UK London, v. 3, n. 3, p. 225–244, 2020. Citado na página 28.
- CHEUNG, P. C. The nutritional and health benefits of mushrooms. **Nutrition Bulletin**, Wiley Online Library, v. 35, n. 4, p. 292–299, 2010. Citado na página 15.

- CHU, L. *et al.* Genome mining as a biotechnological tool for the discovery of novel marine natural products. **Critical reviews in biotechnology**, Taylor & Francis, v. 40, n. 5, p. 571–589, 2020. Citado na página 33.
- CHUGH, R. M. *et al.* Fungal mushrooms: a natural compound with therapeutic applications. v. 13, 2022. Publisher: Frontiers Media SA. Citado na página 21.
- COCK, P. J. *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. Citado na página 38.
- CONSORTIUM, I. P. D. G. *et al.* Imputation of sequence variants for identification of genetic risks for parkinson's disease: a meta-analysis of genome-wide association studies. **The Lancet**, Elsevier, v. 377, n. 9766, p. 641–649, 2011. Citado na página 32.
- CORNISH, T. C.; KRICKA, L. J.; PARK, J. Y. A biopython-based method for comprehensively searching for eponyms in pubmed. **MethodsX**, Elsevier, v. 8, p. 101264, 2021. Citado na página 38.
- COSIC, I.; COSIC, D.; LONCAREVIC, I. Rrm prediction of erythrocyte band3 protein as alternative receptor for sars-cov-2 virus. **Applied Sciences**, MDPI, v. 10, n. 11, p. 4053, 2020. Citado na página 28.
- COUDRAY, N. *et al.* Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. **Nature medicine**, Nature Publishing Group US New York, v. 24, n. 10, p. 1559–1567, 2018. Citado na página 34.
- COZZOLINO, S. M. F. **Biodisponibilidade de nutrientes**. [S.l.]: Editora Manole, 2015. Citado na página 24.
- DUTTA, A. *et al.* An efficient convolutional neural network for coronary heart disease prediction. **Expert Systems with Applications**, Elsevier, v. 159, p. 113408, 2020. Citado na página 33.
- D'AGARO, E. Artificial intelligence used in genome analysis studies. **The EuroBiotech Journal**, v. 2, n. 2, p. 78–88, 2018. Citado na página 13.
- EICHLER, M.; ŞAHİN, G. G.; GUREVYCH, I. Linspector web: A multilingual probing suite for word representations. **arXiv preprint arXiv:1907.11438**, 2019. Citado na página 40.
- ELBAUM, S.; KARRE, S.; ROTHERMEL, G. Improving web application testing with user session data. In: IEEE. **25th International Conference on Software Engineering, 2003. Proceedings**. [S.l.], 2003. p. 49–59. Citado na página 38.
- FALANDYSZ, J.; BOROVÍČKA, J. Macro and trace mineral constituents and radionuclides in mushrooms: health benefits and risks. **Applied microbiology and biotechnology**, Springer, v. 97, p. 477–501, 2013. Citado na página 50.
- FLOREZ, J. C. Mining the genome for therapeutic targets. **Diabetes**, Am Diabetes Assoc, v. 66, n. 7, p. 1770–1778, 2017. Citado na página 32.

- GHOSH, S. K.; GHOSH, A. Classification of gene expression patterns using a novel type-2 fuzzy multigranulation-based SVM model for the recognition of cancer mediating biomarkers. **Neural Comput & Applic**, v. 33, n. 9, p. 4263–4281, maio 2021. ISSN 1433-3058. Disponível em: <https://doi.org/10.1007/s00521-020-05241-7>. Citado na página 47.
- GIL-MARTINS, E. *et al.* Dysfunction of abc transporters at the blood-brain barrier: Role in neurological disorders. **Pharmacology & therapeutics**, Elsevier, v. 213, p. 107554, 2020. Citado na página 29.
- GODSWILL, A. G. *et al.* Health benefits of micronutrients (vitamins and minerals) and their associated deficiency diseases: A systematic review. **International Journal of Food Sciences**, v. 3, n. 1, p. 1–32, 2020. Citado na página 23.
- GOMES, G.; LUDERMIR, T. Otimizacao de pesos e funcoes de ativacao de redes neurais aplicadas na previsao de series temporais. **arXiv preprint arXiv:2107.14370**, 2021. Citado na página 34.
- GOMES, V. C.; QUEIROZ, G. R.; FERREIRA, K. R. An overview of platforms for big earth observation data management and analysis. **Remote Sensing**, MDPI, v. 12, n. 8, p. 1253, 2020. Citado na página 40.
- GONG, W. *et al.* Whole genome sequence of an edible and medicinal mushroom, hericium erinaceus (basidiomycota, fungi). **Genomics**, Elsevier, v. 112, n. 3, p. 2393–2399, 2020. Citado nas páginas 27 e 28.
- GULATI, A. *et al.* Impact of climate change, variability, and extreme rainfall events on agricultural production and food insecurity in orissa. **ISPRS Archives**, v. 38, n. 8, p. W3, 2009. Citado na página 17.
- HALLGREN, J. *et al.* Deeptmhmm predicts alpha and beta transmembrane proteins using deep neural networks. **BioRxiv**, Cold Spring Harbor Laboratory, p. 2022–04, 2022. Citado nas páginas 47, 50 e 52.
- HANNIGAN, G. D. *et al.* A deep learning genome-mining strategy for biosynthetic gene cluster prediction. **Nucleic Acids Research**, v. 47, n. 18, p. e110–e110, out. 2019. ISSN 0305-1048, 1362-4962. Disponível em: <https://academic.oup.com/nar/article/47/18/e110/5545735>. Citado nas páginas 38 e 43.
- HAWKSWORTH, D. L.; LÜCKING, R. Fungal diversity revisited: 2.2 to 3.8 million species. *microbiol. spectr.* 5. **Microbiology Spectr. Am. Soc. Microbiol. Press**, v. 5, p. 1–17, 2017. Citado nas páginas 18 e 19.
- HEWAMALAGE, H.; BERGMEIR, C.; BANDARA, K. Recurrent neural networks for time series forecasting: Current status and future directions. **International Journal of Forecasting**, Elsevier, v. 37, n. 1, p. 388–427, 2021. Citado na página 34.
- HOSSAIN, E.; BABAR, M. A.; PAIK, H.-y. Using scrum in global software development: a systematic literature review. In: **2009 Fourth IEEE International Conference on Global Software Engineering**. [S.l.]: Ieee, 2009. p. 175–184. Citado na página 56.
- HYDE, K. D. The numbers of fungi. **Fungal Diversity**, Springer, v. 114, n. 1, p. 1–1, 2022. Citado na página 18.

- JAIN, S.; SINGH, D. Central dogma of molecular biology and expression of genetic information. **A Textbook of Molecular Biotechnology**, IK International Pvt Ltd, p. 213, 2009. Citado na página 26.
- JIMÉNEZ, E. M. *et al.* The potential of arbuscular mycorrhizal fungi to enhance metallic micronutrient uptake and mitigate food contamination in agriculture: prospects and challenges. **New Phytologist**, Wiley Online Library, v. 242, n. 4, p. 1441–1447, 2024. Citado na página 13.
- JONES, C. **From local to global: engaging in the world's food challenges through a mushroom cultivation case study**. Tese (Doutorado) — Cardiff University, 2021. Citado na página 15.
- KALAČ, P. A review of chemical composition and nutritional value of wild-growing and cultivated mushrooms. **Journal of the Science of Food and Agriculture**, Wiley Online Library, v. 93, n. 2, p. 209–218, 2013. Citado na página 22.
- KARAFFA, L.; FEKETE, E.; KUBICEK, C. P. The role of metal ions in fungal organic acid accumulation. **Microorganisms**, MDPI, v. 9, n. 6, p. 1267, 2021. Citado na página 25.
- KELLER, N. P. Fungal secondary metabolism: regulation, function and drug discovery. **Nature Reviews Microbiology**, Nature Publishing Group UK London, v. 17, n. 3, p. 167–180, 2019. Citado na página 27.
- KHAN, N. S.; ABID, A.; ABID, K. A novel natural language processing (nlp)-based machine translation model for english to pakistan sign language translation. **Cognitive Computation**, Springer, v. 12, p. 748–765, 2020. Citado na página 37.
- KHAN, S. M. *et al.* MU-PseUDep: A deep learning method for prediction of pseudouridine sites. **Computational and Structural Biotechnology Journal**, v. 18, p. 1877–1883, 2020. ISSN 20010370. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S2001037020303421>. Citado na página 44.
- KIM, G. B. *et al.* DeepTFactor: A deep learning-based tool for the prediction of transcription factors. **Proc. Natl. Acad. Sci. U.S.A.**, v. 118, n. 2, p. e2021171118, jan. 2021. ISSN 0027-8424, 1091-6490. Disponível em: <https://pnas.org/doi/full/10.1073/pnas.2021171118>. Citado na página 43.
- KONG, L. *et al.* Mining influential genes based on deep learning. **BMC Bioinformatics**, v. 22, n. 1, p. 27, dez. 2021. ISSN 1471-2105. Disponível em: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-03972-5>. Citado na página 44.
- KOPILER, A. A. *et al.* Redes neurais artificiais e suas aplicações no setor elétrico. **Revista de Engenharias da Faculdade Salesiana**, v. 1, n. 9, p. 27–33, 2019. Citado na página 34.
- KUHAR, F. *et al.* Delimitation of funga as a valid term for the diversity of fungal communities: the fauna, flora & funga proposal (ff&f). **IMA Fungus**, Springer, v. 9, n. 2, p. A71–A74, 2018. Citado na página 18.

- KUMAR, V. *et al.* Role of macronutrient in health. **World Journal of Pharmaceutical Research**, v. 6, n. 3, p. 373–381, 2017. Citado na página 23.
- KUMARI, K. Mushrooms as source of dietary fiber and its medicinal value: A review article. **Journal of Pharmacognosy and Phytochemistry**, AkiNik Publications, v. 9, n. 2, p. 2075–2078, 2020. Citado na página 15.
- KWAN, J. L. *et al.* Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. **Bmj**, British Medical Journal Publishing Group, v. 370, 2020. Citado na página 33.
- LARSON, N. B. *et al.* A clinician's guide to bioinformatics for next-generation sequencing. **Journal of Thoracic Oncology**, Elsevier, 2022. Citado na página 27.
- LEE, W. *et al.* Role of mushrooms in neurodegenerative diseases. **Medicinal Mushrooms: Recent Progress in Research and Development**, Springer, p. 223–249, 2019. Citado na página 22.
- LI, H. *et al.* Reviewing the world's edible mushroom species: A new evidence-based classification system. **Comprehensive Reviews in Food Science and Food Safety**, Wiley Online Library, v. 20, n. 2, p. 1982–2014, 2021. Citado na página 19.
- LI, M. *et al.* Deep-LC: A Novel Deep Learning Method of Identifying Non-Small Cell Lung Cancer-Related Genes. **Front. Oncol.**, v. 12, p. 949546, jul. 2022. ISSN 2234-943X. Disponível em: <https://www.frontiersin.org/articles/10.3389/fonc.2022.949546/full>. Citado na página 44.
- LI, S. *et al.* Whole genome sequence of an edible mushroom stropharia rugosoannulata (daqiugaigu). **Journal of Fungi**, MDPI, v. 8, n. 2, p. 99, 2022. Citado na página 27.
- LIAO, W. *et al.* A review of graph neural networks and their applications in power systems. **Journal of Modern Power Systems and Clean Energy**, SGEPRI, v. 10, n. 2, p. 345–360, 2021. Citado na página 33.
- LIU, D.-M.; DONG, C. Recent advances in nano-carrier immobilized enzymes and their applications. **Process Biochemistry**, Elsevier, v. 92, p. 464–475, 2020. Citado na página 28.
- LIU, Y. *et al.* Application of deep learning algorithm on whole genome sequencing data uncovers structural variants associated with multiple mental disorders in african american patients. **Molecular psychiatry**, Nature Publishing Group UK London, v. 27, n. 3, p. 1469–1478, 2022. Citado na página 32.
- MACIEL, B. K. C. *et al.* Necessidades nutricionais: mudanças com o envelhecimento. In: **Congresso Internacional de Envelhecimento Humano**. [S.l.: s.n.], 2015. Citado na página 17.
- MARTÍ-CARRERAS, J.; MAES, P. Human cytomegalovirus genomics and transcriptomics through the lens of next-generation sequencing: revision and future challenges. **Virus Genes**, Springer, v. 55, n. 2, p. 138–164, 2019. Citado na página 31.
- MATOS¹, W. L. N. *et al.* Aplicação de rede neural artificial na classificação de tipos de íris. 2021. Citado na página 34.

- MCBROOME, J.; TURAKHIA, Y.; CORBETT-DETIG, R. Bte: a python module for pandemic-scale mutation-annotated phylogenetic trees. **Journal of Open Source Software**, v. 7, n. 77, p. 4433, 2022. Citado na página 39.
- MENDES, A. C. d. O. *et al.* Olatecg: ferramenta de bioinformática para o ensino de genética no ensino médio. Universidade Federal de Mato Grosso, 2022. Citado nas páginas 37 e 38.
- MESSINA, A. *et al.* Biograph: a web application and a graph database for querying and analyzing bioinformatics resources. **BMC systems biology**, Springer, v. 12, p. 75–89, 2018. Citado na página 38.
- MOHANTA, T. K.; BAE, H. The diversity of fungal genome. **Biological procedures online**, Springer, v. 17, p. 1–9, 2015. Citado na página 27.
- MONACO, A. *et al.* A primer on machine learning techniques for genomic applications. **Computational and Structural Biotechnology Journal**, v. 19, p. 4345–4359, 2021. ISSN 20010370. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S2001037021003111>. Citado nas páginas 38 e 44.
- NAGAI, R. A.; SBRAGIA, R. As origens da metodologia ágil: de onde saímos e onde estamos? uma revisão sistemática da literatura. v. 14, n. 1, p. 11–41, 2023. Citado na página 55.
- NANDI, S.; SIKDER, R.; ACHARYA, K. Secondary metabolites of mushrooms: A potential source for anticancer therapeutics with translational opportunities. **Advancing Frontiers in Mycology & Mycotechnology: Basic and Applied Aspects of Fungi**, Springer, p. 563–598, 2019. Citado na página 21.
- NIEGO, A. G. *et al.* Macrofungi as a nutraceutical source: Promising bioactive compounds and market value. **Journal of Fungi**, MDPI, v. 7, n. 5, p. 397, 2021. Citado na página 21.
- NUGROHO, A.; KUSUMAWARDANI, S. *et al.* Distributed classifier for sdgs topics in online news using rabbitmq message broker. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [S.l.], 2020. v. 1577, n. 1, p. 012026. Citado nas páginas 40 e 41.
- OLIVATTO, T. F. Identificação automática de rampas de acessibilidade apoiada por visão computacional a partir de imagens panorâmicas street-level. Universidade Federal de São Carlos, 2021. Citado nas páginas 35 e 37.
- OLIVEIRA, E. P. de; SUNDFELD, D. Pa-star-web: web server para obtenção do alinhamento múltiplo ótimo de sequências biológicas. In: SBC. **Anais da X Escola Regional de Informática de Goiás**. [S.l.], 2022. p. 189–192. Citado nas páginas 37 e 38.
- PARKER, J. C.; DUNHAM, P. B. Passive cation transport. In: **Red Blood Cell Membranes**. [S.l.]: CRC Press, 2020. p. 507–561. Citado na página 29.
- PERMYAKOV, E. A. Metal binding proteins. **Encyclopedia**, MDPI, v. 1, n. 1, p. 261–292, 2021. Citado nas páginas 28 e 30.

- PRIHODA, D. *et al.* The application potential of machine learning and genomics for understanding natural product diversity, chemistry, and therapeutic translatability. **Natural Product Reports**, Royal Society of Chemistry, v. 38, n. 6, p. 1100–1108, 2021. Citado nas páginas 32 e 33.
- PRIYADARSHINI, E. *et al.* Metal-fungus interaction: Review on cellular processes underlying heavy metal detoxification and synthesis of metal nanoparticles. **Chemosphere**, Elsevier, v. 274, p. 129976, 2021. Citado na página 25.
- RAPINI, A. **Modernizando a taxonomia**. [S.l.]: SciELO Brasil, 2004. 1–4 p. Citado na página 40.
- Rede PENSSAN. **2º Inquérito Nacional sobre Insegurança Alimentar no Contexto da Pandemia da Covid-19 no Brasil**. 2022. Disponível em: <https://pesquisassan.net.br/2o-inquerito-nacional-sobre-inseguranca-alimentar-no-contexto-da-pandemia-da-covid-19-no-brasil/>. Acesso em: 05 de maio 2024. Citado na página 18.
- REIS, F. S. *et al.* Functional foods based on extracts or compounds derived from mushrooms. **Trends in Food Science & Technology**, Elsevier, v. 66, p. 48–62, 2017. Citado na página 14.
- REKADWAD, B.; GONZALEZ, J. M. New generation dna sequencing (ngs): Mining for genes and the potential of extremophiles. **Microbial Applications Vol. 1: Bioremediation and Bioenergy**, Springer, p. 255–268, 2017. Citado na página 13.
- RIZZO, G. *et al.* A review of mushrooms in human nutrition and health. **Trends in Food Science & Technology**, Elsevier, v. 117, p. 60–73, 2021. Citado na página 15.
- RYU, J. Y.; KIM, H. U.; LEE, S. Y. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. **Proc. Natl. Acad. Sci. U.S.A.**, v. 116, n. 28, p. 13996–14001, jul. 2019. ISSN 0027-8424, 1091-6490. Disponível em: <https://pnas.org/doi/full/10.1073/pnas.1821905116>. Citado na página 43.
- SAINI, S.; SHARMA, K. K. Fungal lignocellulolytic enzymes and lignocellulose: a critical review on their contribution to multiproduct biorefinery and global biofuel research. **International Journal of Biological Macromolecules**, Elsevier, v. 193, p. 2304–2319, 2021. Citado na página 27.
- SAMPAIO, D. B. Predição da evapotranspiração de referência usando rede lstm bidirecional e var+ lstm. Serra, 2022. Citado na página 52.
- SANUMA, O. I. *et al.* Sanöma samakönö sama tökö nii pewö oa wi ï tökö waheta: Ana amopö= enciclopédia dos alimentos yanomami (sanöma): Cogumelos. Hutukara Associação Yanomami e Instituto Socioambiental, 2016. Citado na página 19.
- SAYERS, E. W. *et al.* Genbank. **Nucleic acids research**, Oxford University Press, v. 48, n. D1, p. D84, 2020. Citado na página 40.
- SCHEID, S. S. *et al.* Iron biofortification and availability in the mycelial biomass of edible and medicinal basidiomycetes cultivated in sugarcane molasses. **Scientific reports**, Nature Publishing Group UK London, v. 10, n. 1, p. 12875, 2020. Citado na página 68.

- SERRANO, W. Genetic and deep learning clusters based on neural networks for management decision structures. **Neural Comput & Applic**, v. 32, n. 9, p. 4187–4211, maio 2020. ISSN 1433-3058. Disponível em: <https://doi.org/10.1007/s00521-019-04231-8>. Citado nas páginas 38 e 46.
- SILVA, M. L. A. *et al.* Vulnerabilidade social, fome e pobreza nas regiões norte e nordeste do brasil. **Políticas Públicas, Educ e Divers Uma Compreensão Científica do Real**, p. 1083–105, 2020. Citado na página 18.
- SOLEYMANI, F. *et al.* ProtInteract: A deep learning framework for predicting protein–protein interactions. **Computational and Structural Biotechnology Journal**, v. 21, p. 1324–1348, 2023. ISSN 20010370. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S2001037023000296>. Citado nas páginas 31, 38 e 45.
- SOMESHWAR, D. *et al.* Implementation of virtual assistant with sign language using deep learning and tensorflow. In: IEEE. **2020 second international conference on inventive research in computing applications (ICIRCA)**. [S.l.], 2020. p. 595–600. Citado na página 37.
- SOUZA, E. P. d. *et al.* Aplicações do deep learning para diagnóstico de doenças e identificação de insetos vetores. **Saúde em Debate**, SciELO Brasil, v. 43, p. 147–154, 2020. Citado na página 34.
- SRIVASTAVA, A.; NAIK, A. Big data analysis in bioinformatics. **Advances in Bioinformatics**, Springer, p. 405–429, 2021. Citado na página 31.
- STOIAN, D. *et al.* Value-chain development for rural poverty reduction: A reality check and a warning. International Food Policy Research Institute, 2016. Citado na página 17.
- SU, X. *et al.* Method development for cross-study microbiome data mining: Challenges and opportunities. **Computational and Structural Biotechnology Journal**, v. 18, p. 2075–2080, 2020. ISSN 2001-0370. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2001037020303524>. Citado na página 32.
- TANG, Y.; WU, X. Salient object detection using cascaded convolutional neural networks and adversarial learning. **IEEE Transactions on Multimedia**, IEEE, v. 21, n. 9, p. 2237–2247, 2019. Citado na página 36.
- TAPPIA, P. S.; SHAH, A. K. Role of macronutrients in human health and disease. In: **Biomedical Translational Research: From Disease Diagnosis to Treatment**. [S.l.]: Springer, 2022. p. 477–491. Citado na página 23.
- THU, Z. M. *et al.* Bioactive phytochemical constituents of wild edible mushrooms from southeast asia. **Molecules**, v. 25, n. 8, 2020. ISSN 1420-3049. Disponível em: <https://www.mdpi.com/1420-3049/25/8/1972>. Citado na página 19.
- THU, Z. M. *et al.* Bioactive phytochemical constituents of wild edible mushrooms from southeast asia. v. 25, n. 8, p. 1972, 2020. Publisher: MDPI. Citado na página 20.
- TUSNÁDY, G. E.; DOSZTÁNYI, Z.; SIMON, I. Transmembrane proteins in the protein data bank: identification and classification. **Bioinformatics**, Oxford University Press, v. 20, n. 17, p. 2964–2972, 2004. Citado na página 28.

- UAUY, R. Defining and addressing the nutritional needs of populations. **Public Health Nutrition**, Cambridge University Press, v. 8, n. 6a, p. 773–780, 2005. Citado na página 17.
- Uzma *et al.* Gene encoder: a feature selection technique through unsupervised deep learning-based clustering for large gene expression data. **Neural Comput & Applic**, v. 34, n. 11, p. 8309–8331, jun. 2022. ISSN 1433-3058. Disponível em: <https://doi.org/10.1007/s00521-020-05101-4>. Citado nas páginas 38 e 46.
- VARGAS-ISLA, R.; ISHIKAWA, N. K.; PY-DANIEL, V. Contribuições etnomicológicas dos povos indígenas da amazônia. **Biota Amazônia (Biote Amazonie, Biota Amazonia, Amazonian Biota)**, v. 3, n. 1, p. 58–65, 2013. Citado na página 19.
- VENTURELLA, G. *et al.* Medicinal mushrooms: bioactive compounds, use, and clinical trials. **International journal of molecular sciences**, MDPI, v. 22, n. 2, p. 634, 2021. Citado na página 14.
- VINCENT, A. T. *et al.* Next-generation sequencing (ngs) in the microbiological world: How to make the most of your money. **Journal of microbiological methods**, Elsevier, v. 138, p. 60–71, 2017. Citado na página 13.
- WALDEMARIN, R. C. **Desenvolvimento baseado em modelos de serviços adaptadores para ferramentas de bioinformática**. Tese (Doutorado) — Universidade de São Paulo, 2021. Citado nas páginas 37 e 38.
- WANG, R. *et al.* Convolutional recurrent neural networks for text classification. In: IEEE. **2019 international joint conference on neural networks (IJCNN)**. [S.l.], 2019. p. 1–6. Citado na página 35.
- WHITFORD, D. **Proteins: structure and function**. [S.l.]: John Wiley & Sons, 2013. Citado na página 28.
- WŁODARCZAK, P.; SOAR, J.; ALLY, M. Genome mining using machine learning techniques. In: SPRINGER. **Inclusive Smart Cities and e-Health: 13th International Conference on Smart Homes and Health Telematics, ICOST 2015, Geneva, Switzerland, June 10-12, 2015, Proceedings 13**. [S.l.], 2015. p. 379–384. Citado na página 33.
- WRATTEN, L.; WILM, A.; GÖKE, J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. **Nature methods**, Nature Publishing Group US New York, v. 18, n. 10, p. 1161–1168, 2021. Citado na página 31.
- YANG, G. R.; WANG, X.-J. Artificial neural networks for neuroscientists: a primer. **Neuron**, Elsevier, v. 107, n. 6, p. 1048–1070, 2020. Citado nas páginas 33, 34 e 37.
- YANG, T.-H.; YANG, Y.-C.; TU, K.-C. regCNN: identifying *Drosophila* genome-wide cis-regulatory modules via integrating the local patterns in epigenetic marks and transcription factor binding motifs. **Computational and Structural Biotechnology Journal**, v. 20, p. 296–308, 2022. ISSN 20010370. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S2001037021005249>. Citado nas páginas 38 e 45.

- YEH, S.-J.; YEH, T.-Y.; CHEN, B.-S. Systems drug discovery for diffuse large b cell lymphoma based on pathogenic molecular mechanism via big data mining and deep learning method. **International Journal of Molecular Sciences**, MDPI, v. 23, n. 12, p. 6732, 2022. Citado na página 32.
- YUAN, Y. *et al.* Whole genome sequence of auricularia heimuer (basidiomycota, fungi), the third most important cultivated mushroom worldwide. **Genomics**, Elsevier, v. 111, n. 1, p. 50–58, 2019. Citado na página 27.
- ZHANG, H. *et al.* Predicting lncrna-disease associations using network topological similarity based on deep mining heterogeneous networks. **Mathematical biosciences**, Elsevier, v. 315, p. 108229, 2019. Citado na página 33.
- ZHOU, Z. *et al.* The versatile roles of testrapanins in cancer from intracellular signaling to cell-cell communication: cell membrane proteins without ligands. **Cell & Bioscience**, Springer, v. 13, n. 1, p. 59, 2023. Citado na página 29.
- ZOLTÁN, J. Nyomozás a nyomelemek mentális világában. **Clinical Neuroscience/Idegyogyaszati Szemle**, v. 72, 2019. Citado na página 24.
- ZULKOWER, V.; ROSSER, S. Dna features viewer, a sequence annotation formatting and plotting library for python. **Bioinformatics**, 2020. Citado na página 39.