INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO SUL DE MINAS GERAIS

CÂMPUS MUZAMBINHO

Curso de Ciência da Computação

Fortunato de Figueiredo Roncholeta

Análise de métodos sobre classificação de proteínas com base em inteligência artificial

---

Muzambinho

2023

Fortunato de Figueiredo Roncholeta

Análise de métodos sobre classificação de proteínas com base em inteligência artificial

Trabalho de Conclusão de Curso apresentado ao Curso de Ciência da Computação, do Instituto Federal de Educação Ciência e Tecnologia do Sul de Minas Gerais - Câmpus Muzambinho, como requisito parcial à obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Gustavo Jose da Silva

Muzambinho

2023

COMISSÃO EXAMINADORA

Documento assinado digitalmente
gov.br GUSTAVO JOSE DA SILVA
Data: 19/12/2023 13:29:44-0300
Verifique em https://validar.iti.gov.br

Prof 1.: Gustavo Jose da Silva

(Orientador)

Documento assinado digitalmente
gov.br JOAO MARCELO RIBEIRO
Data: 20/12/2023 17:01:54-0300
Verifique em https://validar.iti.gov.br

Prof 2.: João Marcelo Ribeiro

(Professor convidado)

Documento assinado digitalmente
gov.br RAMON GUSTAVO TEODORO MARQUES DA SIL\
Data: 19/12/2023 16:18:12-0300
Verifique em https://validar.iti.gov.br

Prof 3.: Ramon Gustavo Teodoro Marques da Silva

(Professor convidado)

Muzambinho, ____ de _____de 2023

“Eu acredito que às vezes são as pessoas que ninguém espera nada que fazem as coisas que ninguém consegue imaginar..”
(Alan Turing)

RONCHOLETA, Fortunato de Figueiredo Roncholeta. Análise de métodos sobre classificação de proteínas com base em inteligência artificial. 2023. 6. E: 65f>. Trabalho de Conclusão de Curso (Curso Ciência da Computação) – Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais – Câmpus Muzambinho, Muzambinho, 2023.

RESUMO

As proteínas efetoras desempenham um papel fundamental em processos celulares e no desenvolvimento de medicamentos genéticos, tornando sua classificação uma área de estudo essencial.O estudo baseia-se em pesquisas anteriores conduzidas no Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais (IF Sul de Minas), onde foram aplicados métodos de IA, incluindo redes neurais e algoritmos de classificação. O objetivo principal é avaliar a eficácia dessas abordagens anteriores de classificações de proteínas efetoras.O projeto envolve uma metodologia que abrange revisão bibliográfica, estudo aprofundado das técnicas de IA, análise criteriosa dos resultados obtidos e a elaboração de relatórios. A ênfase será colocada na comparação e classificação dos métodos de acordo com sua precisão.Espera-se que este estudo contribua significativamente para o aprimoramento da classificação de proteínas efetoras, com implicações importantes no campo do desenvolvimento de medicamentos genéticos. Ao investigar a eficácia de métodos de IA, espera-se proporcionar insights valiosos que possam beneficiar futuras pesquisas e aplicações práticas nessa área em constante evolução.

Palavras-chave: Proteínas Efetoras, Inteligência artificial, Classificação ,Análise de Resultados , Redes Neurais, Algoritmos KNN, Medicamentos Genéticos

RONCHOLETA, Fortunato de Figueiredo Roncholeta. Análise de métodos sobre classificação de proteínas com base em inteligência artificial. 2023. 6. E: 65f>. Trabalho de Conclusão de Curso (Curso Ciência da Computação) – Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais – Câmpus Muzambinho, Muzambinho, 2023.

ABSTRACT

Effector proteins play a fundamental role in cellular processes and the development of genetic medicines, making their classification an essential area of study. The study is based on previous research conducted at the Federal Institute of Education, Science, and Technology of Southern Minas Gerais (IF Sul de Minas), where AI methods, including neural networks and classification algorithms, were applied. The main objective is to evaluate the effectiveness of these previous approaches to effector protein classification.

The project involves a methodology that includes a literature review, an in-depth study of AI techniques, a thorough analysis of the results obtained, and the preparation of reports. Emphasis will be placed on the comparison and classification of methods based on their accuracy.

It is expected that this study will contribute significantly to the improvement of effector protein classification, with important implications in the field of genetic medicine development. By investigating the effectiveness of AI methods, valuable insights are expected to be provided that can benefit future research and practical applications in this ever-evolving field.

Keywords: Effector Proteins, Artificial Intelligence, Classification, Results Analysis, Neural Networks, KNN Algorithms, Genetic Medicines.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1. INTRODUÇÃO

1.1 Contextualização e Motivação

Durante décadas, o DNA foi inicialmente percebido como uma estrutura meramente física. No entanto, um avanço crucial surgiu na década de 1940, quando o DNA foi identificado como o possível portador da informação genética, decorrente de estudos sobre hereditariedade em bactérias (ALBERTS et al., 2017). Somente em 1953, os pesquisadores Francis Crick, James Watson e Maurice Wilkins desvendaram a estrutura tridimensional da molécula de DNA. Esse marco impulsionou pesquisas na biologia molecular moderna, abrindo caminho para a concepção de medicamentos com base no código genético individual. Unindo esse conhecimento à compreensão das proteínas, a ideia era desenvolver fármacos mais eficazes, causando menos impacto no indivíduo e minimizando efeitos colaterais.

Ao explorar microrganismos causadores de doenças como tuberculose e tétano, descobriu-se que, ao entrar em contato com o hospedeiro, esses microrganismos secretam proteínas capazes de modificar os processos celulares, inibindo funções como a apoptose. Essas proteínas, denominadas proteínas efetoras, comprometem as células, tornando o sistema de defesa do organismo menos eficaz no combate à infecção. Recentemente, pesquisas focadas em proteínas efetoras revelaram a possibilidade de classificá-las com base em características próprias, como hidropatia e quantidade de aminoácidos polares básicos na região do C-Terminal.

Com o avançar do tempo, diversas investigações foram conduzidas, incorporando técnicas de inteligência artificial para aprimorar a classificação dessas proteínas. Este projeto tem como objetivo avaliar métodos de inteligência artificial e os resultados obtidos em pesquisas anteriores, visando identificar estratégias mais eficientes.

.

1.2 Objetivos

1.2.1 Objetivo Geral

Este estudo tem como objetivo central realizar uma análise detalhada e minuciosa dos métodos de inteligência artificial previamente empregados em pesquisas dedicadas à classificação de proteínas efetoras. O objetivo deste projeto é aprofundar a compreensão e avaliação dos métodos de inteligência artificial aplicados à classificação de proteínas efetoras. Buscamos identificar, destacar e analisar criticamente os resultados mais promissores e eficazes obtidos por meio desses métodos, com foco especial em redes neurais perceptron e algoritmos de classificação, como KNN, SVM e Random Forest. Além disso, pretendemos propor e explorar novas abordagens e ideias inovadoras que possam contribuir significativamente para o avanço dessa área de pesquisa em constante evolução.

1.2.2 Objetivos Específicos

- Avaliar se os métodos de redes neurais perceptron e algoritmos de classificação, como KNN, SVM e Random Forest, proporcionam mais eficácia na identificação de proteínas efetoras, a partir da sequência de nucleotídeos do DNA.
- Avaliar o desempenho dos métodos de inteligência artificial previamente empregados para essa finalidade.
- Realizar uma comparação abrangente dos métodos de classificação estudados.
- Elaborar relatórios detalhados identificando e classificando os métodos mais eficientes com base em critérios específicos.

2. REVISÃO DE LITERATURA

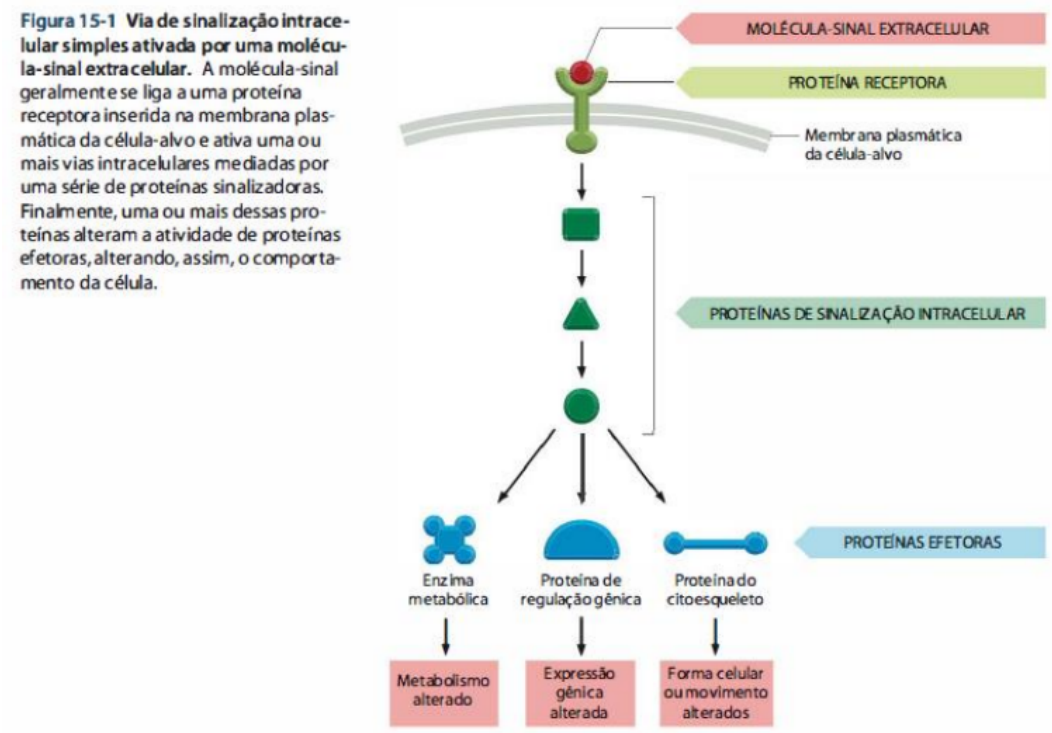
2.1 Proteínas

Proteínas, nada mais são que um conjunto de aminoácidos, que quando unidos entre si por ligações peptídicas recebem o nome de proteínas. Hoje existem um total de 20 tipos de aminoácidos, que são codificados diretamente no DNA eles podem se combinar das mais variadas formas entre si gerando assim diferentes tipos de proteínas. Uma molécula de proteína consiste em uma longa cadeia não ramificada desses aminoácidos, e cada um está ligado aos aminoácidos adjacentes por ligações peptídicas covalentes, podendo ser chamadas também de polipeptídeos. Elas possuem um sequência de aminoácidos exclusiva de cada proteínas, podendo conter milhares em uma única célula. Do ponto de vista químico, as proteínas são as moléculas mais complexas e mais sofisticadas que conhecemos (ALBERTS et al., 2017).

2.2 Proteínas efetoras

Microrganismos patogênicos, como os responsáveis pela tuberculose e o tétano, possuem uma estratégia intrigante para infectar as células do hospedeiro. Eles secretam proteínas efetoras, que atuam dentro das células hospedeiras, alterando seus processos normais. Essas proteínas efetoras têm a capacidade de inibir funções celulares, como a apoptose, tornando a célula do hospedeiro um ambiente mais favorável para a sobrevivência do microrganismo invasor. Isso cria um desafio para o sistema imunológico do corpo, que não consegue eliminar a infecção devido às mudanças induzidas pelas proteínas efetoras. A figura mostra o processo de secreção de uma proteína efetora.

Figura 1:Módulos de sinalização celular



Fonte: Sinalização Celular USP, FMRP-USP, SP, BRASIL

No entanto, pesquisas recentes realizadas por Remédio (2017), têm explorado a possibilidade de classificar essas proteínas efetoras com base em suas características específicas. Algumas dessas características incluem a hidropatia total, a quantidade de aminoácidos polares básicos na região do C-Terminal e a presença de sinais de localização, como nuclear ou mitocondrial. Esses estudos aprofundam nossa compreensão das interações entre microrganismos e células hospedeiras e podem abrir portas para o desenvolvimento de tratamentos mais eficazes contra doenças causadas por esses microrganismos.

2.2.1 Hidropatia Total

Segundo Voet (2013), a hidropatia é capaz de refletir o nível de hidropaticidade, o que significa que pode indicar se uma molécula apresenta características hidrofóbicas, demonstrando repulsa à água, ou hidrofílicas, exibindo afinidade com a molécula de água. A hidropatia é um índice que atribui valores a cada aminoácido, sendo positivo para aminoácidos hidrofóbicos e negativo para aminoácidos hidrofílicos. A soma desses valores revela a hidropatia total da proteína. Conforme descrito por Remédio (2017), a hidropatia representa o grau de hidropaticidade de um aminoácido, revelando suas propriedades de

aversão ou afinidade com a água. Os aminoácidos hidrofóbicos são associados a valores positivos de hidropatia, enquanto os aminoácidos hidrofílicos possuem valores negativos. Kyte e Doolittle desenvolveram uma tabela com base em experimentos para atribuir os valores de hidropatia a cada aminoácido, fornecendo uma referência valiosa para entender as propriedades físico-químicas desses componentes fundamentais da estrutura proteica.

Tabela 1:Escala de hidropatia

Aminoácido	Letra	Valor
Isoleucina	I	4,5
Valina	V	4,2
Leucina	L	3,8
Fenilalanina	F	2,8
Cisteína	C	2,5
Metionina	M	1,9
Alanina	A	1,8
Glicina	G	-0,4
Treonina	T	-0,7
Serina	S	-0,8
Triptofano	W	-0,9
Tirosina	Y	-1,3
Prolina	P	-1,6
Histidina	H	-3,2
Ácido Glutâmico	E	-3,5
Ácido Aspártico	D	-3,5
Glutamina	Q	-3,5
Asparagina	N	-3,5
Lisina	K	-3,9
Arginina	R	-4,5

Fonte: Kyte e Doolittle(1982)

2.2.2 Hidropatia Média

Para calcular a Hidropatia Média de uma proteína, realiza-se uma média aritmética simples dividindo o valor da Hidropatia Total pelo número de aminoácidos contidos na proteína. A obtenção da Hidropatia Média da proteína requer a realização de uma divisão, na qual o valor da Hidropatia Total da proteína é dividido pelo número de aminoácidos que compõem a mesma.

2.2.3 HIdropatia no C-Terminal

De acordo com Remédio (2017), a hidropatia no C-Terminal de uma proteína serve como um indicador de suas propriedades hidrofílicas ou hidrofóbicas nessa região específica. Proteínas que demonstram uma forte afinidade com a molécula de H2O, ou seja, que possuem características hidrofílicas no C-Terminal, têm uma probabilidade maior de serem identificadas como proteínas efetoras, conforme observado por Lockwood et al. (2011).

A hidropatia no C-Terminal é o local onde se encontram proteínas com características hidrofílicas, e é nessa região que pode ser identificado o sinal de translocação da mesma proteína efetora, conforme documentado por Lockwood et al. (2011)

2.2.4 Carga no C-Terminal

Com base no trabalho de Nelson e Cox (2011), a classificação dos aminoácidos pode ser realizada de três maneiras distintas, levando em consideração a carga elétrica, a estrutura e o tamanho. A categorização por carga elétrica envolve a identificação de aminoácidos com grupos R que apresentam cargas positivas, sendo designados como aminoácidos polares básicos, e aqueles com cargas negativas, que são conhecidos como aminoácidos polares ácidos. A determinação da carga no C-Terminal é obtida por meio do cálculo da diferença entre o número de aminoácidos polares básicos e o número de aminoácidos polares ácidos, considerando a atribuição de carga +1 para os aminoácidos polares básicos e carga -1 para os aminoácidos polares ácidos antes da realização do cálculo.

2.2.5 Aminoácidos Polares Básicos no C-Terminal

Meyer et al. (2013) afirmam que a presença de no mínimo três aminoácidos polares básicos no C-Terminal é uma característica comum em todas as proteínas efetoras identificadas. Esses aminoácidos desempenham um papel-chave na identificação das proteínas efetoras e são essenciais para o sucesso da predição.

2.2.6 Sinal de Localização Nuclear

Com base nas contribuições de Nguyen Ba et al. (2009) e Meyer et al. (2013), o núcleo celular é dotado de um mecanismo de seleção que utiliza poros para determinar quais materiais podem acessá-lo. O Sinal de Localização Nuclear consiste em uma sequência específica de aminoácidos que serve como um guia para direcionar a proteína até o núcleo, seu propósito é fornecer à proteína as instruções necessárias para navegar até o núcleo. Essas sequências de aminoácidos podem ser identificadas por meio de algoritmos computacionais

que reconhecem esses padrões. Existem duas formas principais de sinal de localização Nuclear: uma monopartida, que segue o padrão PKKKRKV, e outra bipartida, caracterizada por sequências KR separadas por espaços variados. As proteínas que têm como destino o núcleo possuem essas sequências de sinal de localização nuclear. Elas interagem com os receptores e complexos de poros, garantindo que a proteína alcance o núcleo sem obstáculos.

2.2.7 Sinal de Localização Mitocondrial

Com base em Meyer et al. (2013), proteínas contêm uma sequência denominada Sinal de Localização Mitocondrial, que guia sua entrada na mitocôndria. Essa sequência se divide em duas partes ao interagir com a mitocôndria. Ela é caracterizada por conter principalmente os aminoácidos Arginina, Leucina, Serina e Alanina, com pelo menos dois deles com carga positiva. Essa sequência pode formar uma dupla hélice e não possui acidez. Essas sequências estão localizadas no N-terminal da proteína e na região adjacente a ele. Elas variam em comprimento e na sequência de aminoácidos, mas geralmente incluem R (Arginina), L (Leucina), S (Serina) e A (Alanina), com pelo menos dois aminoácidos com carga positiva. Essas sequências são essenciais para a translocação das proteínas nas mitocôndrias, conforme destacado por Claros e Vincens (1995).

2.2.8 Domínio Espiral Enrolada

Em proteínas efetoras, é possível identificar sequências que apresentam semelhanças com características de espécies eucariontes. Isso ocorre porque os patógenos frequentemente imitam o comportamento do hospedeiro. No entanto, ao analisar proteínas efetoras, é fundamental considerar a presença de domínios específicos, como o domínio espiral enrolada, conforme observado por Lockwood et al. (2011), de acordo com Silva (2015), o domínio espiral enrolada, conhecido como "dupla hélice", possui uma estrutura que se assemelha a uma corda e é comumente encontrada em proteínas de interação, principalmente em eucariotos.

As proteínas em espiral dupla, também chamadas de "Coiled Coil", compartilham a característica de ter duas ou mais hélices alfa torcidas ao redor de uma outra hélice. Essas espirais duplas são notavelmente prevalentes em estruturas proteicas, podendo variar em número de duas a sete hélices, conforme destacado por Trigg (2011).

2.2.9 Domínio Prenilação

Conforme destacado por Silva (2015), a estabilidade das proteínas está vinculada a uma alteração permanente que ocorre após a tradução. Certas bactérias, como mencionado por Ivanov et al. (2010), exploram a prenilação do hospedeiro para facilitar o trajeto das proteínas efetoras em direção às organelas associadas à membrana durante a infecção intracelular.

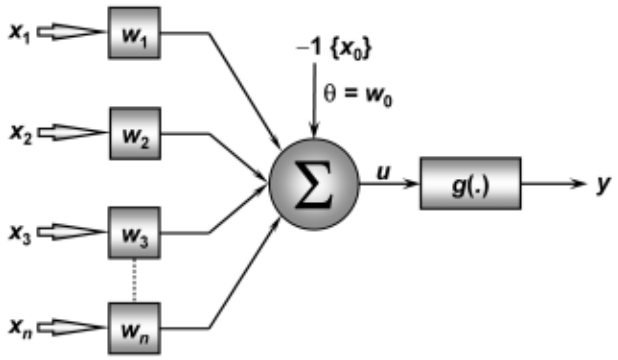
2.3 Algoritmos de Classificação

Com a popularização das inteligências artificiais muitos métodos de classificação estão sendo criados, tais métodos vêm mostrando bons resultados com o passar do tempo. Técnicas como por exemplo o machine learning (ou, em português, Aprendizado de máquina) que utiliza combinação de algoritmos para realizar análises de grandes volumes de dados e criar padrões que estabelecem conexões, aprendendo a executar tarefas de forma inteligente sem a intervenção do homem, já vem sendo usadas em diversas áreas atualmente. Com estas técnicas seria possível dado uma entrada de características de um determinado alvo, classificá-lo em grupos específicos que compartilhem de uma ou mais características.

2.3.1 Perceptron

O perceptron é uma das formas mais elementares de rede neural, empregado para classificar padrões que podem ser separados linearmente, isto é, quando esses padrões podem ser divididos em dois grupos distintos por uma linha ou hiperplano. Basicamente, um perceptron é constituído por um único neurônio que possui pesos sinápticos ajustáveis e um termo de viés (bias). O algoritmo usado para ajustar esses parâmetros livres na rede neural foi inicialmente desenvolvido por Frank Rosenblatt nas décadas de 1950 e 1960 em seu modelo conceitual do perceptron como uma analogia ao funcionamento cerebral (HAYKIN, et al., 2011).

Figura 2:Modelo Rede perceptron

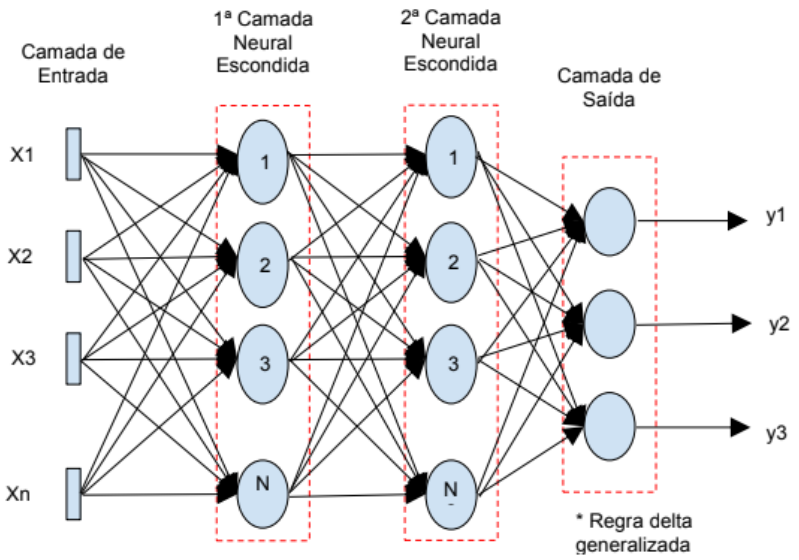


Fonte: Redes Neurais Artificiais para engenharia e ciências aplicadas. 2010. p 58

2.3.2 MLP

Dentre as diversas categorias de Redes Neurais Artificiais (RNAs), uma notável é o Multilayer Perceptron, frequentemente considerado um avanço em relação ao Perceptron tradicional. Este modelo, originalmente desenvolvido por Rosenblatt, destaca-se pela sua capacidade de incluir múltiplas camadas intermediárias de neurônios (HAYKIN, et al., 2011).

Figura 3:Modelo de perceptron multicamadas



Fonte: Braga,2007,p . 49

O Multilayer Perceptron é altamente flexível e eficaz em lidar com tarefas complexas, graças à capacidade de usar várias camadas intermediárias, cada uma com múltiplos neurônios, o que aprimora seu poder de processamento e aprendizado. Isso possibilita a execução de tarefas avançadas e a extração eficiente de informações dos dados de entrada.

2.3.3 KNN

O algoritmo K-nearest neighbors (KNN) é uma ferramenta versátil comumente empregada para tarefas de classificação e regressão em análise de dados. De acordo com Peterson (2009), Ele se destaca quando possuímos conhecimento limitado sobre a distribuição dos dados. No contexto da classificação, o KNN utiliza métodos de cálculo de distância, como a distância euclidiana, conforme apresentado na equação.

$$d(X_i, X_l) = \sqrt{(x_{i1} - x_{(l1)})^2 + (x_{i2} - x_{(l2)})^2 + . \quad . \quad . + (x_{in} - x_{(ln)})^2}$$

Essa equação quantifica a distância entre os dados ao calcular a raiz quadrada da soma dos quadrados das diferenças entre as características dos dados comparados.

Uma vez que as distâncias foram calculadas, o KNN classifica novos pontos de dados com base na proximidade dos K vizinhos mais próximos. Em resumo, o KNN opera em duas etapas fundamentais: identificação dos vizinhos mais próximos e determinação da classe à qual pertence um novo dado, levando em consideração um número específico de vizinhos. Esse algoritmo representa uma ferramenta inestimável ao lidar com dados de estrutura desconhecida, como indicado por Peterson (2009) e Cunningham e Delany (2007).

2.3.4 Random Forest

O algoritmo Random Forest foi desenvolvido com a árvore de decisão como sua base subjacente. A abordagem, conforme delineada por Breiman (2001), envolve o uso de várias árvores de decisão para votar e selecionar a classe mais predominante. De acordo com Amaratunga, Cabrera e Lee (2008), o algoritmo Random Forest demonstra uma eficácia notável quando cada classificador individual exibe um desempenho sólido. Além disso, a diversidade entre as árvores e sua conexão limitada são fatores essenciais para o sucesso do Random Forest.

2.3.5 SVM

As Máquinas de Vetores de Suporte (SVM, Support Vector Machines) são um tipo de algoritmo de classificação supervisionada projetado para separar eficientemente dois grupos distintos. Como Lorena e Carvalho (2003) explicam, as SVMs têm suas raízes na fundamentação teórica das Teorias Estatísticas.

2.4 Trabalhos Relacionados

REMÉDIO (2017) Durante sua pesquisa, alcançamos uma notável taxa de precisão de 97,5% após treinar a rede neural com 100 proteínas efetoras e 100 proteínas não efetoras, considerando oito características essenciais. Observamos que a diminuição da precisão pós-treinamento está relacionada principalmente aos atributos de sinal de localização mitocondrial e sinal de localização nuclear, que, em algumas situações, não permitem uma distinção clara entre proteínas efetoras e não efetoras.

OLIVEIRA (2018) em sua pesquisa intitulada, *Sistema de predição de proteínas efetoras baseado na análise comparativa de algoritmos de classificação de inteligência artificial ,* utilizou uma base de dados compostas por 249 proteínas efetoras e 249 proteínas não efetoras contendo proteínas com 9 atributos para a classificação sendo elas, hidropatia total, hidropatia média, hidropatia no C-Terminal, carga no C-Terminal, aminoácidos polares básicos no C-Terminal, sinal de localização nuclear, sinal de localização mitocondrial, domínio espiral enrolada e domínio prenilação. durante a pesquisa ele obteve utilizando redes neurais de 1 camada oculta com 5 neurônios uma acurácia de 86,2%.

COELHO (2019) realizou uma pesquisa tentando aprimorar métodos de inteligência artificial para a predição de proteínas efetoras. Durante o trabalho foi dividido o banco de dados uma parte com treinamento e a outra parte para o pós treinamento , Foram selecionadas 180 proteínas efetoras e 180 proteínas não efetoras para compor a base de dados de treinamento. Já para o pós treinamento foi usado o restante das proteínas, 69 proteínas efetoras e 69 proteínas não efetoras. O método utilizado foi o algoritmo de KNN baseando-se em 9 atributos selecionados que foram hidropatia total, hidropatia média, hidropatia no C Terminal, carga no C-Terminal, quantidade de aminoácidos polares básicos no C-Terminal, sinal de localização nuclear, sinal de localização mitocondrial, domínio espiral enrolada e domínio prenilação. Como resultado utilizando o método do KNN (K-Nearest Neighbors), com 7 vizinhos e a Métrica de Minkowski, obteve-se uma acurácia de 88.40% .

3. METODOLOGIA

Este Trabalho constitui uma investigação minuciosa das pesquisas anteriores conduzidas no Instituto Federal - Campus Sul de Minas Muzambinho, com uma ênfase particular na tarefa desafiadora da classificação de proteínas efetoras. Este estudo adota uma abordagem que recorre a sofisticadas técnicas de inteligência artificial para a análise das proteínas efetoras, representando um avanço significativo na busca por soluções mais eficazes neste domínio científico.

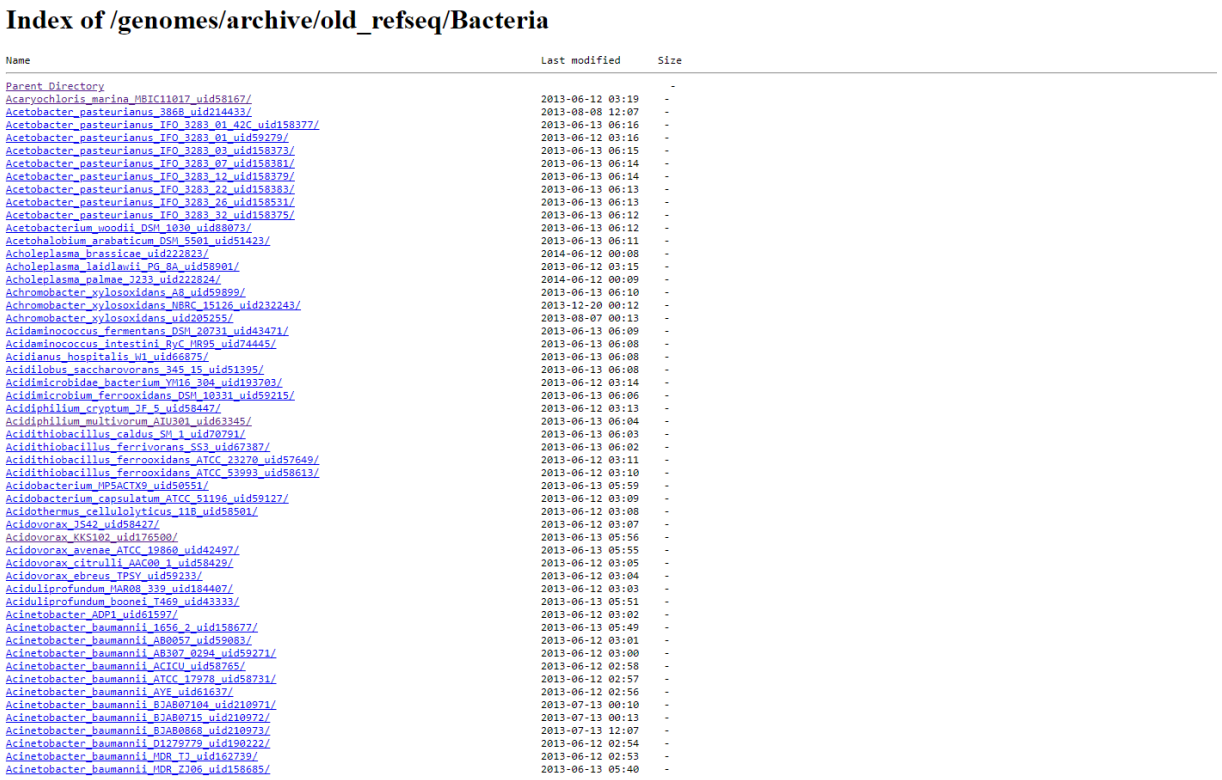
3.1 Análise

Durante a fase de Análise, foram examinadas e compiladas informações essenciais para a condução do projeto. Este estudo abrangeu a revisão do referencial teórico e das pesquisas anteriores realizadas no Instituto Campus Muzambinho, no período de 2016 a 2019. Além disso, foram investigados os resultados anteriores, juntamente com as metodologias empregadas, para embasar de maneira sólida o desenvolvimento deste trabalho.

3.2 Base de dados

Os dados para o projeto foram adquiridos da plataforma NCBI (*National Center for Biotechnology Information*), uma compilação abrangente de registros genômicos, transcritos e proteicos, conhecida por sua integração e anotações detalhadas. Especificamente, as sequências de proteínas foram obtidas através do acesso ao endereço https://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/ via ftp que corresponde a (*File Transfer Protocol*) um protocolo que viabiliza a transferência de arquivos entre computadores em redes, incluindo a internet. Ele desempenha um papel fundamental na eficiente e segura movimentação de dados de um ponto para outro.

Figura 4: Página do portal NCBI



Index of /genomes/archive/old_refseq/Bacteria

Name	Last modified	Size
Parent Directory		-
Acaryochloris_marina_MBIC11017_uid58167/	2013-06-12 03:19	-
Acetobacter_pasteurianus_386B_uid214433/	2013-08-08 12:07	-
Acetobacter_pasteurianus_IFO_3283_01_42C_uid158377/	2013-06-13 06:16	-
Acetobacter_pasteurianus_IFO_3283_01_uid59279/	2013-06-12 03:16	-
Acetobacter_pasteurianus_IFO_3283_03_uid158373/	2013-06-13 06:15	-
Acetobacter_pasteurianus_IFO_3283_07_uid158381/	2013-06-13 06:14	-
Acetobacter_pasteurianus_IFO_3283_12_uid158379/	2013-06-13 06:14	-
Acetobacter_pasteurianus_IFO_3283_22_uid158383/	2013-06-13 06:13	-
Acetobacter_pasteurianus_IFO_3283_26_uid158531/	2013-06-13 06:13	-
Acetobacter_pasteurianus_IFO_3283_32_uid158375/	2013-06-13 06:12	-
Acetobacterium_woodii_DSM_1030_uid88073/	2013-06-13 06:12	-
Acetohalobium_arabaticum_DSM_5501_uid51423/	2013-06-13 06:11	-
Acholeplasma_brassicae_uid222823/	2014-06-12 00:08	-
Acholeplasma_laidlawii_PG_8A_uid58901/	2013-06-12 03:15	-
Acholeplasma_palmae_J233_uid222824/	2014-06-12 00:09	-
Achromobacter_xylosoxidans_A8_uid59899/	2013-06-13 06:10	-
Achromobacter_xylosoxidans_NBRC_15126_uid232243/	2013-12-20 00:12	-
Achromobacter_xylosoxidans_uid205255/	2013-08-07 00:13	-
Acidaminococcus_fermentans_DSM_20731_uid43471/	2013-06-13 06:09	-
Acidaminococcus_intestini_RyC_MR95_uid74445/	2013-06-13 06:08	-
Acidianus_hospitalis_W1_uid66875/	2013-06-13 06:08	-
Acidilobus_saccharovorans_345_15_uid51395/	2013-06-13 06:08	-
Acidimicrobidae_bacterium_YM16_304_uid193703/	2013-06-12 03:14	-
Acidimicrobium_ferrooxidans_DSM_10331_uid59215/	2013-06-13 06:06	-
Acidiphilium_cryptum_JF_5_uid58447/	2013-06-12 03:13	-
Acidiphilium_multivorum_AIU301_uid63345/	2013-06-13 06:04	-
Acidithiobacillus_caldus_SM_1_uid70791/	2013-06-13 06:03	-
Acidithiobacillus_ferrivorans_SS3_uid67387/	2013-06-13 06:02	-
Acidithiobacillus_ferrooxidans_ATCC_23270_uid57649/	2013-06-12 03:11	-
Acidithiobacillus_ferrooxidans_ATCC_53993_uid58613/	2013-06-12 03:10	-
Acidobacterium_MP5ACTX9_uid50551/	2013-06-13 05:59	-
Acidobacterium_capsulatum_ATCC_51196_uid59127/	2013-06-12 03:09	-
Acidothermus_cellulolyticus_11B_uid58501/	2013-06-12 03:08	-
Acidovorax_JS42_uid58427/	2013-06-12 03:07	-
Acidovorax_KKS102_uid176500/	2013-06-13 05:56	-
Acidovorax_avenae_ATCC_19860_uid42497/	2013-06-13 05:55	-
Acidovorax_citrulli_AAC00_1_uid58429/	2013-06-12 03:05	-
Acidovorax_ebreus_TPSY_uid59233/	2013-06-12 03:04	-
Aciduliprofundum_MAR08_339_uid184407/	2013-06-12 03:03	-
Aciduliprofundum_boonei_T469_uid43333/	2013-06-13 05:51	-
Acinetobacter_ADP1_uid61597/	2013-06-12 03:02	-
Acinetobacter_baumannii_1656_2_uid158677/	2013-06-13 05:49	-
Acinetobacter_baumannii_AB0057_uid59083/	2013-06-12 03:01	-
Acinetobacter_baumannii_AB307_0294_uid59271/	2013-06-12 03:00	-
Acinetobacter_baumannii_ACICU_uid58765/	2013-06-12 02:58	-
Acinetobacter_baumannii_ATCC_17978_uid58731/	2013-06-12 02:57	-
Acinetobacter_baumannii_AYE_uid61637/	2013-06-12 02:56	-
Acinetobacter_baumannii_BJAB07104_uid210971/	2013-07-13 00:10	-
Acinetobacter_baumannii_BJAB0715_uid210972/	2013-07-13 00:13	-
Acinetobacter_baumannii_BJAB0868_uid210973/	2013-07-13 12:07	-
Acinetobacter_baumannii_D1279779_uid190222/	2013-06-12 02:54	-
Acinetobacter_baumannii_MDR_TJ_uid162739/	2013-06-12 02:53	-
Acinetobacter_baumannii_MDR_ZJ06_uid158685/	2013-06-13 05:40	-

Fonte: Do Autor (2023)

A estrutura do arquivo FASTA, preferencialmente adotada pelo NCBI para armazenar informações sobre proteínas, é delineada da seguinte maneira: inicia com um cabeçalho que fornece detalhes sobre as características específicas da proteína, seguido pela representação integral da sequência de aminoácidos que a compõem. Essa formatação facilita a organização e recuperação eficiente de dados genéticos.

Quadro 1: Arquivo FASTA

```
>gi|73666638|ref|YP_302654.1| chaperone protein DnaJ [Ehrlichia canis str. Jake]
MSKSDYYELLGVSKNATSEEIKKAYRKMALKYHPDTNPGNKEAEEKFKELSEAYDVLIDQDKRAAYDKYG
HNAFDGAAGRGGFDFNSGFSGDFSDIFNDLFGGGFGSRGGRGSSRRSEGAAGSDLRFDVEITLEDSFNGK
KVPISYVTYVKCSSCSGSGSEGSAKSVQCNTCHGAGSVRTQQGFFTIERTCHVCNGEGEIIQNKCKKCSG
SGRVRDEVNLLVTIPKGIESGNKIRLNGKGEAGYRGARSGDLYVYSNIQKHKFFTRSGPDLYCTVPIKMT
LAALGGHIEMPSIDGTWTKVKVPEGSQSGDKLRLKEKGMPVINSSKRGDMYIQITVETPVKLTKKQKELL
QKFDDEPNVDCNPQSTGFFQKVKSFWKDIRSN
```

Fonte: Do Autor (2023)

A representação gráfica acima ilustra uma amostra de sequência de aminoácidos pertencente a uma proteína específica do tipo *Ehrlichia canis*. A linha inicial em destaque traz a identificação da sequência, incluindo nome e tipo, enquanto as linhas subsequentes apresentam a sequência propriamente dita dos aminoácidos da proteína, formatada de acordo com o padrão fasta. Essa notação facilita a interpretação e o manuseio eficaz de informações genéticas.

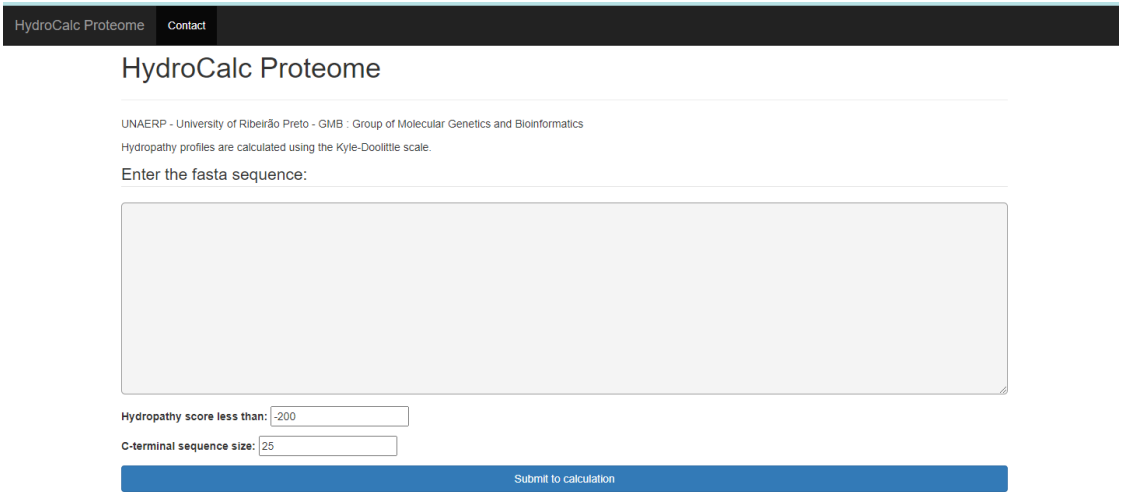
Dessa forma, em todos os trabalhos, incluindo este, realizou-se a coleta de dados de diversas sequências de aminoácidos de proteínas, como exemplificado anteriormente. Esses dados foram submetidos a uma análise e separados em dois documentos distintos, um destinado a proteínas efetoras e outro a não efetoras. Após essa segmentação, procedeu-se à análise dessas proteínas nas ferramentas de bioinformática, tais como *Hydrocalc Proteome*, *NLStradamus*, *TargetP* e *Coiled Coil Prediction*, cujas descrições serão apresentadas no próximo tópico. Essas ferramentas foram empregadas para extrair características que contribuirão nas etapas subsequentes, como a classificação no software WEKA e a implementação em Python.

3.3 Ferramentas

Nesta seção, serão abordadas as ferramentas fundamentais utilizadas na pesquisa, tais como *Hydrocalc Proteome, NLStradamus, TargetP, Coiled Coil Prediction*, e a linguagem de programação Python.

O *Hydrocalc Proteome* é uma aplicação desenvolvida em HTML e JavaScript com o objetivo de avaliar a hidropatia de proteínas.

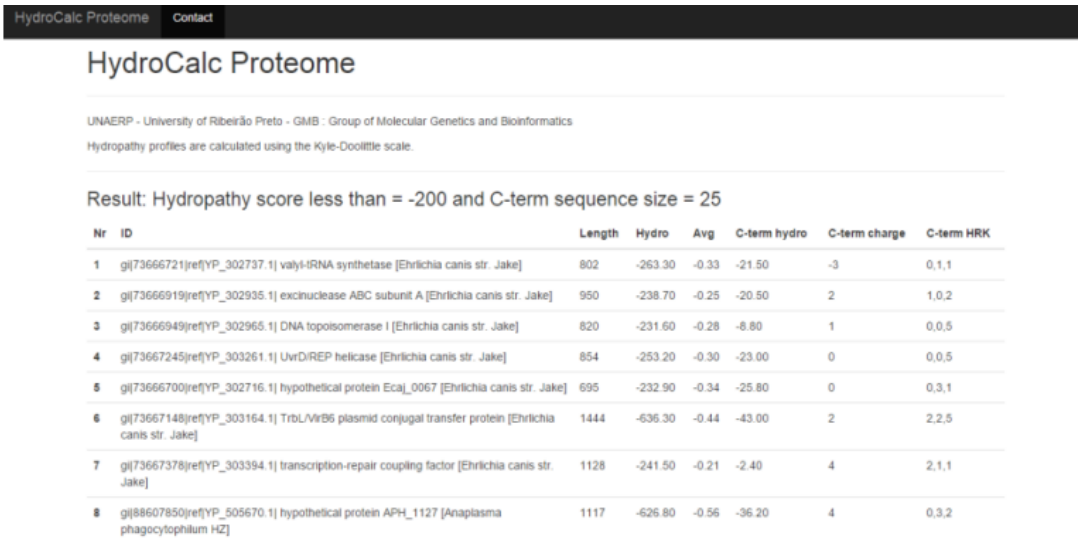
Figura 5: HydroCalc Proteome Interface



Fonte: Do Autor (2023)

A ferramenta extrai cinco características das proteínas, incluindo hidropatia total, hidropatia média, hidropatia do C-terminal, carga do C-terminal e aminoácidos polares básicos no C-terminal. Após a extração das características das proteínas, o próximo passo envolve a criação dos conjuntos de dados de teste e treino. Estes conjuntos iniciais são formados com base nessas características, preparando-se assim a base para as etapas subsequentes de análise e classificação.

Figura 6: HydroCalc Proteome (Classificação)



Fonte: Do Autor (2023)

Após a obtenção dos dados com o *Hydrocalc Proteome*, a pesquisa conduzida por Oliveira (2016) propôs a inclusão de mais características adicionais nos dados de avaliação. Nas pesquisas subsequentes, às características sinal de localização nuclear, sinal de localização Mitocondrial, e a estrutura de espiral enrolada foram incorporadas aos dados para uma análise mais abrangente. A ferramenta *NLStradamus* foi utilizada para examinar o sinal de localização nuclear em cada proteína, abrangendo sequências de proteínas efetoras e não efetoras. Além disso, a ferramenta *TargetP* desempenhou um papel significativo na identificação de padrões relacionados ao Sinal de Localização Mitocondrial. E por fim a ferramenta *Coiled Coil Prediction*, desenvolvida no polo de Bioinformática de Lyon, foi aplicada para identificar a estrutura de espiral enrolada. Essa ferramenta é crucial para a análise da presença desse domínio específico, sendo relevante para a compreensão das características estruturais das proteínas estudadas.

Coelho (2019) incorporou avanços à pesquisa ao seguir as sugestões dos estudos anteriores, introduzindo uma nova característica: o domínio de prenilação. Essa inclusão representa uma evolução na análise, aproveitando as descobertas prévias para aprimorar ainda mais a compreensão das proteínas estudadas.

Desenvolvido pela Universidade de Waikato em 1997, o WEKA é um software Java disponibilizado sob a licença GPL. Ele oferece uma interface gráfica para interação e visualização de dados, com foco em Inteligência Artificial, especialmente em mineração de dados. Entre seus métodos está a Classificação, utilizado por Redes Neurais de Múltiplas Camadas (*Multilayer Perceptron*) para categorizar proteínas.

Figura 7: Software Weka

Fonte: Do Autor (2023)

O WEKA é uma ferramenta abrangente com algoritmos de aprendizado de máquina para mineração de dados, podendo ser aplicados diretamente ou chamados por código Java. O *download* do *Software* pode ser feito em https://www.cs.waikato.ac.nz/ml/weka/. A ferramenta WEKA foi empregada por Oliveira (2016), Remédio (2017) e Oliveira (2018), utilizada para realizar a classificação de proteínas efetoras.

3.4 Python

Python, é uma linguagem de programação reconhecida por sua clareza e versatilidade, desempenhou um papel central nesta pesquisa. Amplamente adotada devido à sua legibilidade e comunidade ativa, Python foi a escolha principal para implementação dos métodos estudados. As bibliotecas *NumPy*, focada em computação científica, e *Scikit-learn*, dedicada à inteligência artificial, desempenharam um papel crucial na manipulação e análise eficiente de dados. Essas Bibliotecas foram utilizadas por Oliveira (2018) e Coelho (2019) com o propósito de comparar métodos entre o *software Weka* e implementações em Python. Os algoritmos KNN (Vizinhos Mais Próximos), *Random Forest* e Máquina de Vetores de Suporte, todos provenientes da biblioteca *Scikit-learn*, foram empregados nesta análise comparativa.

4. RESULTADOS E DISCUSSÃO

Inicialmente, realizou-se um estudo abrangente das pesquisas anteriores conduzidas no Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais (IF Sul de Minas). Durante esse processo, foram compilados os dados mais relevantes de cada pesquisa, incluindo informações sobre a base de dados utilizada, os métodos empregados, as tecnologias envolvidas, a quantidade de atributos analisados e os resultados finais obtidos. Além disso, foram consideradas as conclusões e sugestões finais apresentadas em cada pesquisa.

Oliveira(2016) em sua pesquisa utilizando as 5 principais características, hidropatia total, hidropatia média, hidropatia do C-terminal, carga do C-terminal e aminoácidos polares básicos no C-terminal, que foram classificadas com ajuda do Hydrocalc Proteome utilizando redes neurais perceptron de 6 Neurônios obteve com 5000 épocas a taxa de acerto de 88% de acerto no treinamento e em seguida 87,1134 % nos dados de pós- treinamento. O estudo abrangeu 145 proteínas efetoras e 249 proteínas não efetoras, evidenciando a eficácia dessas características na distinção entre esses dois grupos e verificado que as redes neurais perceptron podem classificar proteínas efetoras com base em características de hidropatia. Como perspectiva de avanço, sugeriu-se explorar a inclusão de mais características em conjunto com as já analisadas e a aplicação de outros tipos de rede neural.

Figura 8: Resultados obtidos por Claudinei de Oliveira (2016)

```
=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances         169               87.1134 %
Incorrectly Classified Instances        25               12.8866 %
Kappa statistic                          0.6668
Mean absolute error                      0.2189
Root mean squared error                  0.3665
Relative absolute error                 43.7842 %
Root relative squared error             73.3053 %
Total Number of Instances              194

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.844     0.121      0.679     0.844     0.752      0.903    yes
                 0.879     0.156      0.949     0.879     0.913      0.903    no
Weighted Avg.    0.871     0.147      0.886     0.871     0.876      0.903

=== Confusion Matrix ===

   a   b   <-- classified as
  38   7 |   a = yes
  18 131 |   b = no
```

Fonte:Utilização de redes neurais perceptron para classificação de proteínas efetoras com base em suas características de hidropatia (2016)

Em 2017, na pesquisa conduzida por Remédio, foram incorporadas três novas características aos dados de análise: sinal de localização nuclear, localização mitocondrial e estrutura espiral enrolada. O treinamento da rede neural, composta por uma camada de 5 neurônios, atingiu uma taxa de acerto de 88% utilizando 5 atributos , 91% utilizando 7 atributos e 97,5% utilizando 8 atributos , enquanto no Pós- treinamento 80% utilizando 5 atributos , 71,9% utilizando 7 atributos e 80,7 % utilizando 8 atributos como mostra a tabela criada por Ricardo V. C. Remédio .

Quadro 2: Resultados obtidos por Ricardo V. C. Remédio (2017)

	5 Atributos	7 Atributos	8 Atributos
Treinamento	88%	91%	97.5%
Pós-Treinamento	80%	72.9%	80.7%

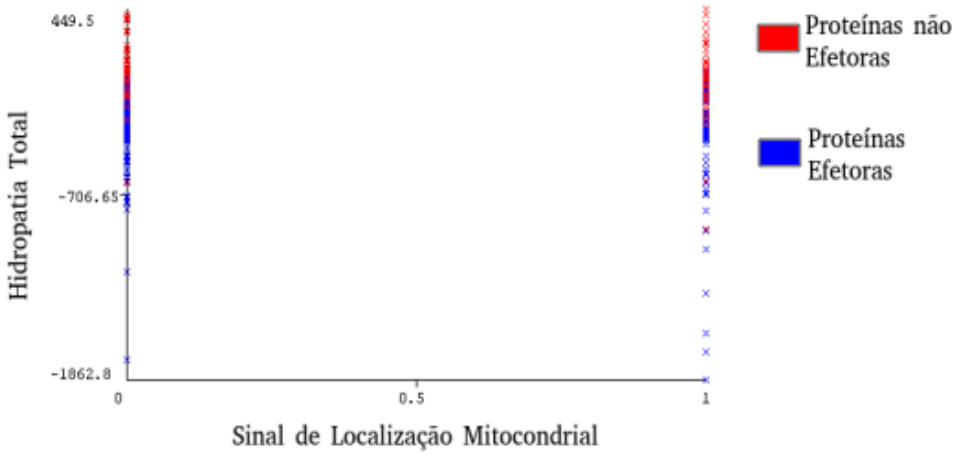
Fonte:Utilização de redes neurais Multilayer Perceptron para classificação de proteínas efetoras (2017)

A inclusão dessas características adicionais nos resultados de treinamento destacou um aumento nas taxas de acerto da rede neural. No entanto, ao examinar os resultados pós-treinamento com 5, 6 e 8 características, observou-se que a adição de 2 atributos resultou em uma queda de 6,1% na taxa de acerto. Surpreendentemente, na última etapa, houve um aumento de 7,8%. A redução no pós-treinamento é atribuída aos atributos de sinal de localização mitocondrial e sinal de localização nuclear, pois essas características fornecem dados que não conseguem discernir efetivamente entre proteínas efetoras e não efetoras. Consequentemente, a rede neural pode enfrentar desafios na convergência para resultados satisfatórios devido a esses atributos específicos.

Em 2018, Gabriel Bianchin de Oliveira conduziu uma pesquisa utilizando o software Weka, desta vez empregando uma base de dados composta por 249 proteínas efetoras e 249 não efetoras. Nos testes realizados com uma única camada e 5 neurônios, foram utilizados 8 atributos: hidropatia total, hidropatia média, hidropatia no C-Terminal, carga no C-Terminal, quantidade de aminoácidos polares básicos no C-Terminal, sinal de localização nuclear, sinal de localização mitocondrial e domínio de espiral enrolada. Os resultados alcançaram uma taxa de acerto de 80,1%, próxima aos obtidos na pesquisa anterior, que atingiu 80,7%.

Posteriormente, foram realizadas análises nas redes neurais artificiais implementadas em Python, utilizando as bibliotecas NumPy e Scikit-learn, com os algoritmos KNN, Random Forest e SVM. O algoritmo que obteve os melhores resultados foi o KNN, alcançando 87,6% em todos os casos, utilizando 7 vizinhos. Ao concluir a pesquisa, Oliveira (2018) destacou a possibilidade de alguns atributos estarem interferindo na classificação. Ele gerou gráficos que ilustram a relação entre os atributos, como o que demonstra a relação entre sinal de localização mitocondrial e hidropatia total, evidenciando a dificuldade em criar uma reta que separa os dados desses dois atributos.

Figura 9: Resultados obtidos por Gabriel Bianchin de Oliveira (2018)



Fonte :Sistema de predição de proteínas efetoras baseado na análise comparativa de algoritmos de classificação de inteligência artificial (2018)

Posteriormente, foi feita uma revisão dos testes, eliminando atributos que poderiam impactar a classificação. O KNN novamente se destacou, alcançando 87,6% em todos os casos, com 7 vizinhos. Na conclusão, comparou os resultados entre Python (NumPy e Scikit-learn) e Weka, observando semelhança nos índices, com o KNN liderando no pós-treinamento. Como sugestão para futuras pesquisas, foi proposto ampliar a base de dados e explorar algoritmos de aprendizado semi-supervisionado.

Na pesquisa conduzida por Hércules de Lima Coelho em 2019, utilizando uma base de dados composta por 249 proteínas, o ponto de partida foi o algoritmo KNN, empregando métricas de distância como Manhattan, Euclidiana e Minkowski. Essa escolha foi motivada pelos bons resultados obtidos na pesquisa anterior com o KNN. Além disso, uma nova característica foi introduzida, totalizando 9 atributos, com a adição do atributo de domínio de prenilação.

Tabela 2: Métrica de Manhattan

Quantidade de Vizinhos Mais Próximos	Acurácia
3	79.71%
5	81.15%
7	86.23%

Fonte :Otimização de Agentes inteligentes para predição de proteínas efetoras (2019)

Tabela 3: Métrica Euclidiana

Quantidade de Vizinhos Mais Próximos	Acurácia
3	79.71%
5	84.78%
7	87.68%

Fonte :Otimização de Agentes inteligentes para predição de proteínas efetoras (2019)

Tabela 4:Métrica de Minkowski

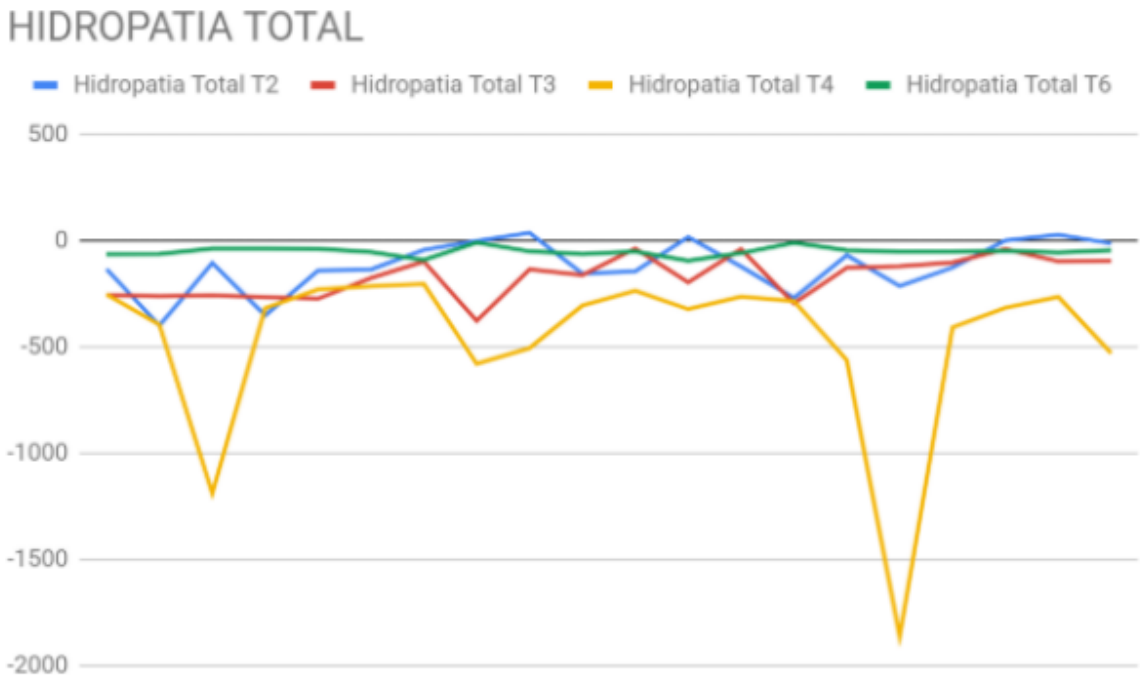
Quantidade de Vizinhos Mais Próximos	Acurácia
3	80.43%
5	83.33%
7	88.40%

Fonte :Otimização de Agentes inteligentes para predição de proteínas efetoras (2019)

Os resultados da pesquisa revelaram as taxas de acerto alcançadas pelo algoritmo ao utilizar as métricas de distância Manhattan, Euclidiana e Minkowski, considerando três, cinco e sete vizinhos para os testes. A métrica de Minkowski apresentou o resultado mais favorável, atingindo uma taxa de acerto de 88,40%.

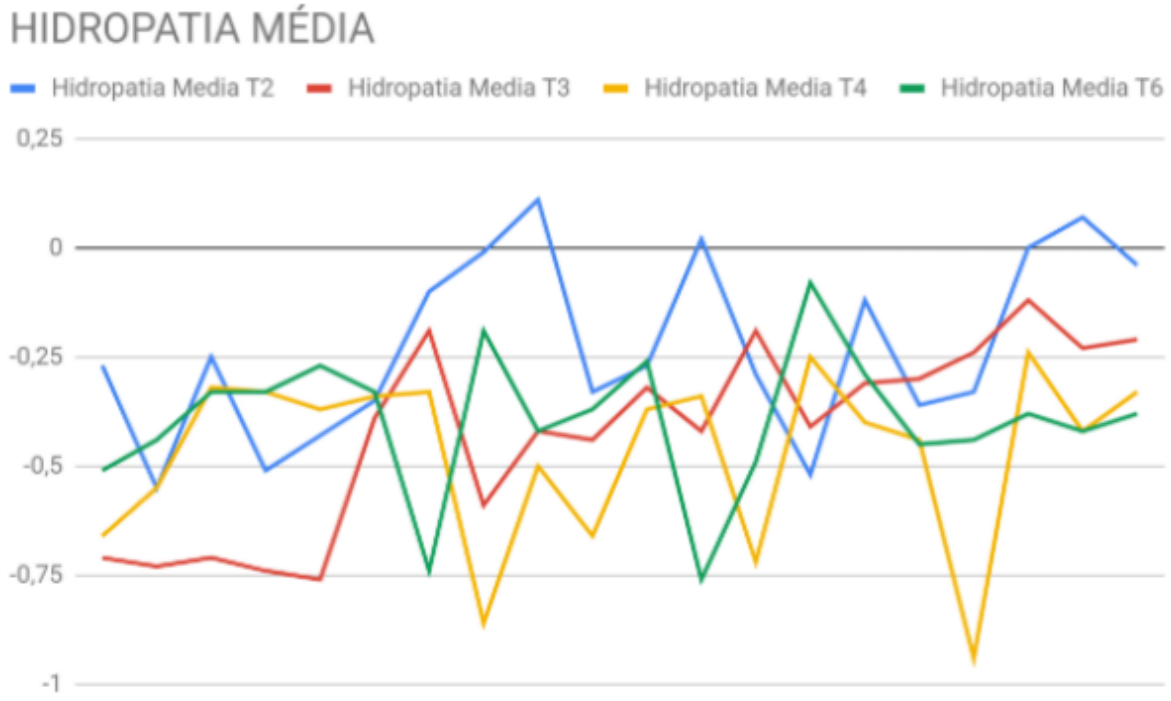
Posteriormente, foi realizada uma investigação para examinar os diferentes sistemas de secreção pelos quais as proteínas são liberadas. O objetivo central deste estudo foi comparar as características das proteínas secretadas nos sistemas dos tipos II, III, IV e VI. Utilizando as nove características mencionadas anteriormente, foram desenvolvidos gráficos para visualizar e comparar os padrões observados em cada conjunto de proteínas associado a cada sistema de secreção.

Figura 10:Hidropatia total das proteínas efetoras.



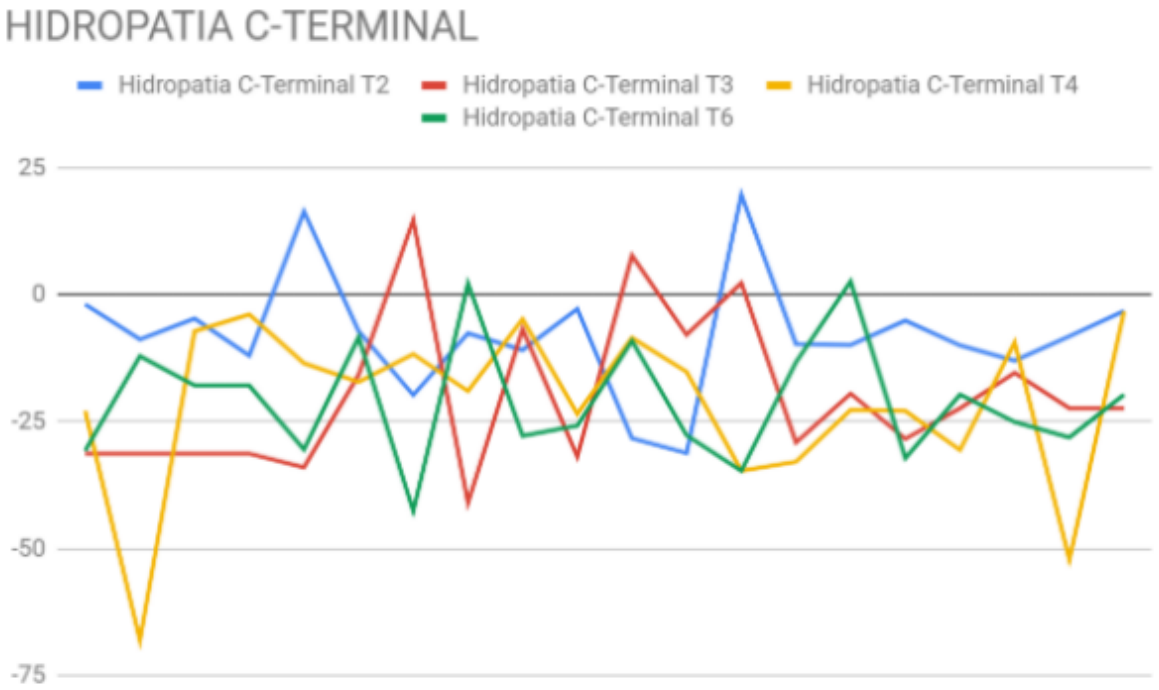
Fonte :Otimização de Agentes inteligentes para predição de proteínas efetoras (2019)

Figura 11:Hidropatia média das proteínas efetoras.



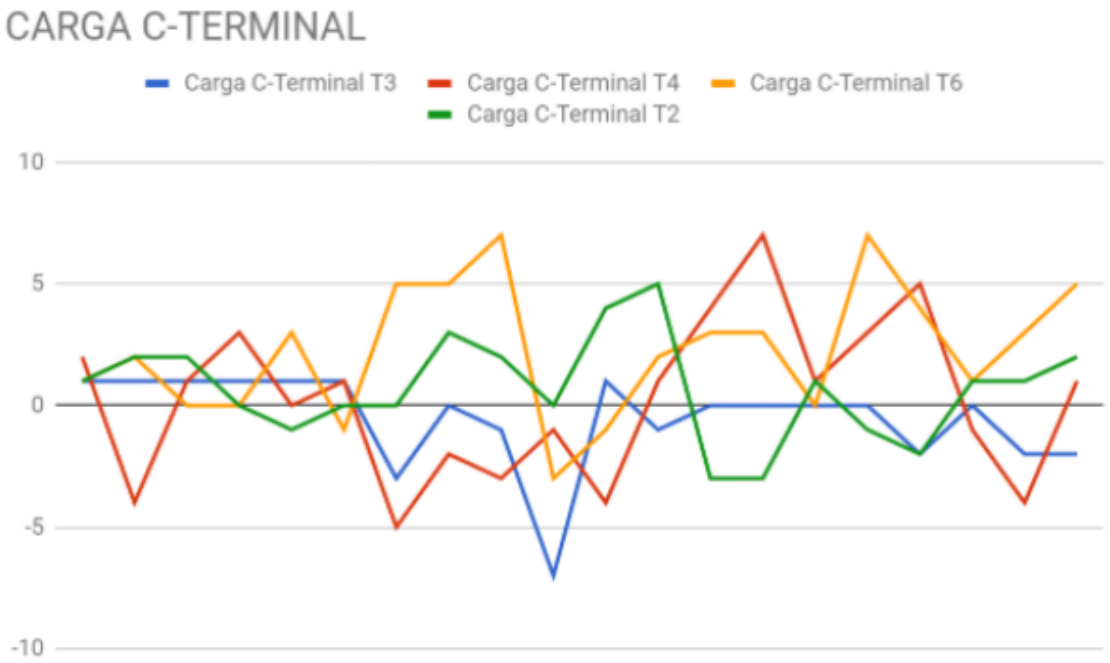
Fonte :Otimização de Agentes inteligentes para predição de proteínas efetoras (2019)

Figura 12:Hidropatia no c-terminal das proteínas efetoras.



Fonte :Otimização de Agentes inteligentes para predição de proteínas efetoras (2019)

Figura 13:Carga no c-terminal das proteínas efetoras.

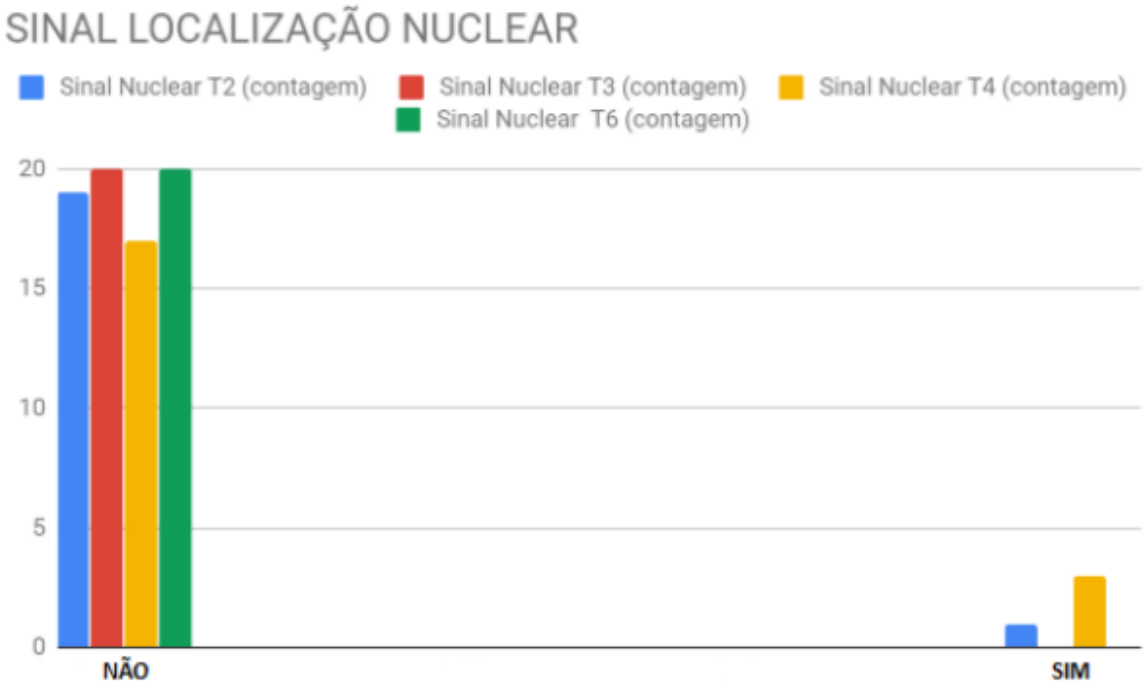


Fonte :Otimização de Agentes inteligentes para predição de proteínas efetoras (2019)

A análise dos gráficos revela padrões distintos em relação às características das proteínas secretadas pelos diferentes tipos de sistemas de secreção. Notavelmente, o atributo de hidropatia total exibe um grau significativo de negatividade nas proteínas secretadas pelo tipo IV, enquanto a hidropatia média revela uma variação considerável, influenciada pela quantidade de aminoácidos que compõem a proteína. Proteínas secretadas pelo tipo II tendem a ser mais hidrofóbicas, sugerido pela variação significativa do atributo de hidropatia média.

As comparações do C-Terminal também indicam alta variação entre proteínas, enquanto a carga no C-Terminal mostra-se mais negativa em proteínas secretadas pelos sistemas de tipo III e IV. Para complementar a análise, foram realizadas comparações gráficas em uma escala binária, distinguindo se a proteína possui ou não a característica em questão.

Figura 14:Presença de sinal de localização nuclear nas proteínas efetoras.

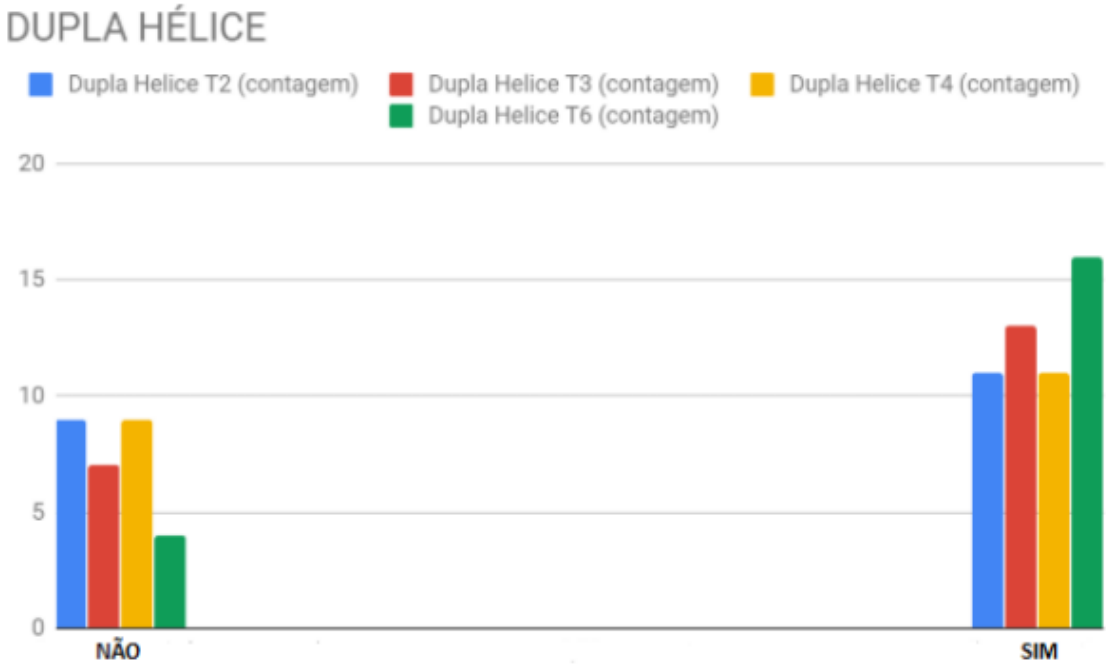


Fonte :Otimização de Agentes inteligentes para predição de proteínas efetoras (2019)

Figura 15:Presença de sinal de localização mitocondrial nas proteínas efetoras.

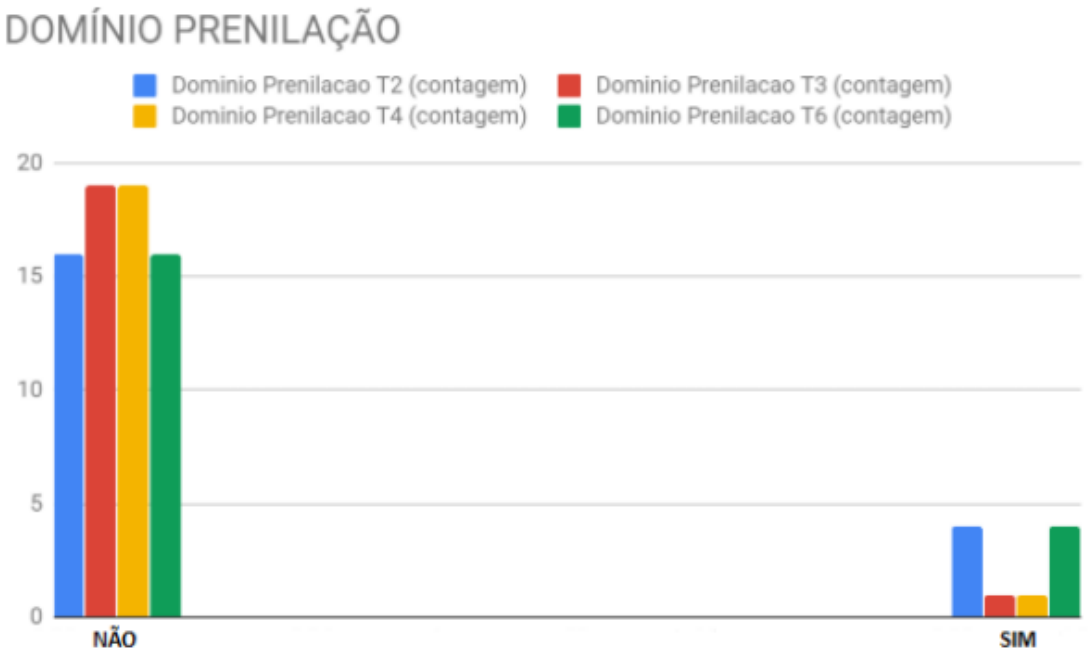
Fonte :Otimização de Agentes inteligentes para predição de proteínas efetoras (2019)

Figura 16:Presença de domínio espiral enrolada/dupla hélice nas proteínas efetoras.



Fonte :Otimização de Agentes inteligentes para predição de proteínas efetoras (2019)

Figura 17:Presença de domínio da prenilação nas proteínas efetoras.



Fonte :Otimização de Agentes inteligentes para predição de proteínas efetoras (2019)

Com os resultados é possível extrair informações cruciais sobre as características das proteínas secretadas pelos diferentes sistemas de secreção. Nota-se que o Sinal de Localização Nuclear teve baixa presença na maioria dos tipos de proteínas efetoras estudadas, enquanto o Sinal de Localização Mitocondrial ocorre frequentemente em proteínas secretadas pelos quatro tipos de sistemas, destacando-se nos tipos II e VI. Em contrapartida, o Domínio Prenilação apresentou baixa incidência, especialmente nas proteínas secretadas pelos sistemas tipo III e IV.

Os dados obtidos apontam que o método de secreção das proteínas exerce impacto nas classificações, e características como Sinal de Localização Mitocondrial e Domínio Espiral Enrolada que podem apresentar variações na presença ou ausência em proteínas efetoras em determinados tipos de sistemas, o que reduz a confiabilidade desses atributos sem levar em conta os tipos mencionados. Em contrapartida, características como Localização Nuclear e Domínio Prenilação revelam diferenças notáveis, indicando serem critérios robustos durante o processo de classificação para os sistemas excretores estudados.

Após a coleta dos resultados, foi desenvolvido um quadro para proporcionar uma visualização mais clara e compreensiva dos dados obtidos. Essa estratégia visa facilitar a análise e interpretação dos resultados, permitindo uma compreensão mais abrangente das tendências e padrões identificados durante a pesquisa.

Quadro 4 : Padrões identificados durante a pesquisa.

Análise				
	Método de IA	**Atributos**	**Dataset**	**Acurácia**
Oliveira (2016)	Redes neurais Perceptron 6 Neuronios	5	145 Efetoras 249 N/ Efetoras	**6 Neurônio - 87,1%**
Remédio (2017)	Redes neurais Perceptron Multi camadas	5, 7 e 8	145 Efetoras 249 N/ Efetoras	5 Atributos - 80% 7 Atributos - 72,9% **8 Atributos - 80,7%**
Oliveira (2016)	KNN Random Forest SVM Rede Neural	9	249 Efetoras 249 N/ Efetoras	Rede neural - 86,2% **KNN - 87,6%** Random-Forest - 83,3% **SVM - 87,6%**
Coelho (2019)	KNN- Manhattan. KNN- Euclidiana. KNN- Minkowski	9	249 Efetoras 249 N/ Efetoras	KNN-Manhattan - 86 % KNN- Euclidiana - 87% KNN- **Minkowski -88%**

Fonte : Do Autor

4.1 Aprofundamento

Após uma análise comparativa dos resultados, tornou-se evidente que os comportamentos biológicos desempenham um papel crucial na classificação de proteínas. Em resposta a essa constatação, um estudo mais aprofundado sobre os mecanismos de secreção de proteínas foi conduzido, considerando a potencial influência desses mecanismos nos resultados obtidos.

Os sistemas de secreção são essenciais para as bactérias, permitindo o transporte de macromoléculas através de membranas sem comprometer sua integridade. Processos fundamentais, como virulência, colonização e mobilidade, dependem da eficiente secreção de moléculas efetoras no ambiente celular imediato e, em alguns casos, no citoplasma do hospedeiro. Como discutido por Ricardi (2013), Jacob (2009) e, mais recentemente, por Chou (2022) em suas pesquisas, os tipos de sistema de secreção podem variar significativamente, dependendo da natureza específica da proteína em questão.Os Sistemas de Secreção Tipo II (T2SSs) são fundamentais em transportando proteínas do periplasma para o meio extracelular. O canal T2SS na membrana externa requer que as proteínas sejam direcionadas ao periplasma pelas vias de secreção Sec ou Tat. Já os Sistemas de Secreção Tipo III (T3SSs), são notáveis por sua estrutura que lembra "agulhas e seringas". Esses sistemas secretam diversos substratos proteicos, atravessando as membranas internas e externas das bactérias. Os Sistemas de Secreção Tipo IV (T4SSs), relacionados à conjugação bacteriana de DNA, podem secretar

diversos substratos em diferentes células-alvo, enquanto o Sistema de Secreção Tipo VI (T6SS) utiliza efetores que se associam à estrutura perfurante de duas maneiras distintas. Os efetores "carga" interagem não covalentemente com um dos componentes da estrutura perfurante, enquanto os efetores "especializados" representam homólogos adicionais deste componente, carregando um domínio efetor adicional fundido covalentemente ao domínio central, geralmente no C-terminal

5. CONSIDERAÇÕES FINAIS

Ao final do estudo foi possível reunir todas as informações das pesquisas anteriores e identificar qual o melhor caminho para se seguir na pesquisa buscando a classificação de Proteínas efetoras com a ajuda de algoritmos de classificação e redes neurais . Percebeu-se que tecnologias como o Python associadas aos algoritmos de Classificação como o KNN podem contribuir muito para o processo como mostrado ao longo das pesquisas, utilizando Métrica de Minkowski onde tivemos uma taxa de acerto de 88,40%, o que se mostra bem relevante . Foi verificado também que algumas características utilizadas como critério de classificação podem estar atrapalhando o processo e que os sistemas de secreções pode influenciar bastante nos resultados

Como pesquisas futuras seria interessante refazer os testes , separando a base de dados por tipo de sistema de secreção retirando as características menos relevantes em cada tipo. Fica a ideia também testar outros tipos de estruturas de algoritmos.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ALBERTS, Bruce et al. **Biologia Molecular da Célula** .6. ED. [S. 1]: Artmed Editora, 2017.

ROCHA, Rafael Silva. **Sinalização Celular**. FMRP-USP,SP,Brasil, 2020.

REMÉDIO, Ricardo Vasconcellos de Carvalho. **Utilização de Redes Neurais Multilayer Perceptron para classificação de proteínas efetoras**. 2017. 44 f. — Trabalho de Conclusão de Curso (Curso Ciência da Computação) – Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais – Câmpus Muzambinho, 2017, Muzambinho.

VOET, Donald; VOET, Judith G. **Bioquímica**. [S.l.]: Artmed Editora, 2013.

LOCKWOOD, Svetlana et al. **Identification of anaplasma marginale type iv secretion system effector proteins.** PLoS One, Public Library of Science, v. 6, n. 11, p. e27724, 2011.

NELSON, D.L.; COX, M.M. **Princípios de Bioquímica de Lehninger**. Porto Alegre: Artmed, 2014, 1273 p.

MEYER, Damien F et al. Searching algorithm for type iv secretion system effectors 1.0: a tool for predicting type iv effectors and exploring their genomic context. **Nucleic acids research**, Oxford University Press, v. 41, n. 20, p. 9218–9229, 2013.

NGUYEN, A.N. et al**. NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction**. BMC Bioinformatics. London, p. 1-11. jun. 2009.

SILVA, Gustavo José da. Identificação e caracterização computacional de proteínas efetoras de bactérias da família Anaplasmataceae. 2015. 108 f. Tese (Doutorado) — Curso de Biotecnologia, Unaerp, Ribeirão Preto, 2015.

TRIGG, J. et al. **Multicoil2: Predicting Coiled Coils and their oligomerization states from sequence in the twilight zone**. Plos One, São Francisco, p. 1-11. ago. 2011.

IVANOV, S. S. et al. **Lipidation by the host prenyltransferase machinery facilitates membrane localization of Legionella pneumophila effector protein**s. Journal of Biological Chemistry. Baltimore, p. 34686-34698. nov. 2010.

HAYKIN, S. **Redes Neurais: Princípios e prática**. [S.l.]: Porto Alegre: Bookman, 2001. 900 p.

SILVA, Ivan Nunes da. **Redes Neurais Artificiais Para Engenharia e Ciências Aplicadas. Fundamentos Teóricos e Aspectos Práticos ,** Sao Paulo , 2010.

BRAGA, A.P.; CARVALHO, A.P.L.F.; LUDERMIR, T.B. **Redes Neurais Artificiais - Teoria e Aplicações**. Rio de Janeiro: LTC, 2007. 140 p.

PETERSON, Leif E. **K-nearest neighbor**. Scholarpedia, v. 4, n. 2, p. 1883, 2009.

CUNNINGHAM, Padraig; DELANY, Sarah Jane. k-nearest neighbour classifiers. **Multiple Classifier Systems**, Springer New York, NY, USA, v. 34, n. 8, p. 1–17, 2007.

AMARATUNGA, Dhammika; CABRERA, Javier; LEE, Yung-Seop. **Enriched random forests**. Bioinformatics, Oxford University Press, v. 24, n. 18, p. 2010–2014, 2008.

LORENA, Ana Carolina; CARVALHO, ACPLF. **Introdução as máquinas de vetores suporte**. Relatório Técnico do Instituto de Ciências Matemáticas e de Computaçao (USP/Sao Carlos), v. 192, 2003.

REMÉDIO, Ricardo Vasconcellos de Carvalho. **Utilização de Redes Neurais Multilayer Perceptron para classificação de proteínas efetoras**. 2017. 44 f. — Trabalho de Conclusão de Curso (Curso Ciência da Computação) – Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais – Câmpus Muzambinho, 2017, Muzambinho.

OLIVEIRA, Gabriel Bianchin de. **Sistema de Predição de Proteínas Efetoras baseado na Análise Comparativa de Algoritmos de Classificação de Inteligência Artificial**. 2018. 44 f. — Trabalho de Conclusão de Curso (Curso Ciência da Computação) – Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais – Câmpus Muzambinho, 2018, Muzambinho.

OLIVEIRA, Claudinei de Oliveira.**Utilização de redes neurais perceptron para classificação de proteínas efetoras com base em suas características de hidropatia**. 2016. 58 f. — Trabalho de Conclusão de Curso (Curso Ciência da Computação) – Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais – Câmpus Muzambinho, 2016, Muzambinho.

COELHO, Hércules de Lima. **Otimização de agente inteligente para predição de proteínas efetoras** . 2019. 36 f. — Trabalho de Conclusão de Curso (Curso Ciência da Computação) – Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais – Câmpus Muzambinho, 2019, Muzambinho.

RICARDI, Ligia Maria Piassi. **Indentificação de proteínas secretadas por duas especies de leptospira , uma patogênica e uma saprofita.** 2013 31 f. – Tese de Pós- Graduação em Biotecnologia USP/Instituto Butantan.

JACOB,Tiago Rinaldi . **Análise da expressão dos genes pertencentes aos sistemas secretórios Tipo III e Tipo IV e genes pthAs em Xanthomonas citri subsp. citri sob condições infectante e não infectante.** 2009. 91 f. Trabalho de Conclusão de Curso Universidade Estadual Paulista Julio de Mesquita Filho.

Chou, L., Lin, YC., Haryono, M. et al. **Modular evolution of secretion systems and virulence plasmids in a bacterial species complex**. BMC Biol 20, 16 (2022). https://doi.org/10.1186/s12915-021-01221-y.

Green Erin. R, Mecsas Joan. **Bacterial Secretion Systems: An Overview.** ASM Journals (2016).