Utilização de Rede Neural Para Predição de Proteínas de Bactérias Secretadas Por Vias Não Clássicas

Luiz Gustavo de Sousa Oliveira

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2020

Luiz Gustavo de Sousa Oliveira

Utilização de Rede Neural Para Predição de Proteínas de Bactérias Secretadas Por Vias Não Clássicas

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Prof. Dr. Anderson Rodrigues dos Santos
Coorientador: Dr. Maria Camila Nardini Barioni

Uberlândia
2020

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

O48
2020

Oliveira, Luiz Gustavo de Sousa, 1992-
Utilização de Rede Neural Para Predição de Proteínas de Bactérias Secretadas Por Vias Não Clássicas [recurso eletrônico] / Luiz Gustavo de Sousa Oliveira. - 2020.

Orientador: Anderson Rodrigues dos Santos.
Coorientadora: Maria Camila Nardini Barioni.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.di.2021.34
Inclui bibliografia.

1. Computação. I. Santos, Anderson Rodrigues dos, 1971-, (Orient.). II. Barioni, Maria Camila Nardini, -, (Coorient.). III. Universidade Federal de Uberlândia. Pós-graduação em Ciência da Computação. IV. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:

Gizele Cristine Nunes do Couto - CRB6/2091

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Ciência da Computação
Av. João Naves de Ávila, nº 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Mestrado Acadêmico, 39 /2020, PPGCO				
Data:	21 de dezembro de 2020	Hora de início:	16:30	Hora de encerramento:	18:35
Matrícula do Discente:	11812CCP021				
Nome do Discente	Luiz Gustavo de Sousa Oliveira				
Título do Trabalho:	Predição de Proteínas de Bactérias Secretadas Por Vias Não Clássicas				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Inteligência Artificial				
Projeto de Pesquisa de vinculação:	-				

Reuniu-se, por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Paulo Henrique Ribeiro Gabriel - FACOM/UFU; José Lopes de Siqueira Neto - DCC/UFMG e Anderson Rodrigues dos Santos - FACOM/UFU, orientador do candidato.

Os examinadores participaram desde as seguintes localidades: José Lopes de Siqueira Neto - Belo Horizonte/MG; Paulo Henrique Ribeiro Gabriel e Anderson Rodrigues dos Santos - Uberlândia/MG. O discente participou da cidade de Uberlândia/MG.

Iniciando os trabalhos o presidente da mesa, Prof. Dr. Anderson Rodrigues dos Santos, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação do Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado.

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Paulo Henrique Ribeiro Gabriel**, **Professor(a) do Magistério Superior**, em 22/12/2020, às 10:57, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Anderson Rodrigues dos Santos**, **Professor(a) do Magistério Superior**, em 30/12/2020, às 15:06, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **José Lopes de Siqueira Neto**, **Usuário Externo**, em 31/12/2020, às 14:05, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2465658** e o código CRC **415E5180**.

Referência: Processo nº 23117.075940/2020-01 SEI nº 2465658

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Os abaixo assinados, por meio deste, certificam que leram e recomendam para a Faculdade de Computação a aceitação da dissertação intitulada "**Utilização de Rede Neural Para Predição de Proteínas de Bactérias Secretadas Por Vias Não Clássicas"** por **Luiz Gustavo de Sousa Oliveira** como parte dos requisitos exigidos para a obtenção do título de Mestre em Ciência da Computação.

Uberlândia, ___ de _____ de _____

Orientador:_____
Prof. Dr. Anderson Rodrigues dos Santos
Universidade Federal de Uberlândia

Coorientador: _____
Prof. Dr. Maria Camila Nardini Barioni
Universidade Federal de Uberlândia

Banca Examinadora:

_____
Prof. Dr. Paulo Henrique Ribeiro Gabriel
Universidade Federal de Uberlândia

_____
Prof. Dr. José Lopes de Siqueira Neto
Universidade Federal de Minas Gerais

Este trabalho é dedicado a todos que tentam contribuir para a melhoria da sociedade através das suas pesquisas científicas.

Agradecimentos

Agradeço a Deus por me proporcionar oportunidades, saúde e conhecimento; ao meu professor orientador Dr. Anderson Rodrigues dos Santos pelas contribuições e à Universidade Federal de Uberlândia por fornecer os recursos que viabilizaram a conclusão do trabalho. Agradeço também a Kelly Nataly Cunha Silva pelo apoio e incentivo na realização deste mestrado.

"Nós só podemos ver um pouco do futuro, mas o suficiente para perceber que há muito a fazer."

(Alan Turing)

Resumo

Apresentamos uma metodologia de predição de proteínas advindas de secreção bacteriana não clássica utilizando redes neurais artificiais. Nosso objetivo é contribuir para a elaboração de vacinas e diagnósticos de doenças a partir de proteínas bacterianas secretada por vias não clássicas. Para esse objetivo, compilamos uma lista de proteínas bacterianas conhecidas advindas de organismos procariotos secretadas pelas vias não clássicas. Essas proteínas foram catalogadas permitindo a criação de um conjunto de treinamento e validação da rede neural artificial. Realizamos uma pesquisa bibliográfica para identificar prováveis descritores e características sinalizadoras deste tipo de secreção bacteriana. Elaboramos uma rede neural supervisionada pelo software *WEKA*. Diversos modelos foram treinados a fim de determinar o melhor grupo de características para a predição de proteínas. Avaliamos o método proposto com a predição de proteínas que não foram utilizadas no grupo de treinamento e comparando com outros dois preditores estudados em literatura correlata, PeNGaRoo (ZHANG et al., 2020) e SecretomeP 2.0 (BENDTSEN et al., 2005). Consideramos nossos resultados satisfatórios, pois apresentaram uma rede neural com acurácia média de 93%. Nosso modelo preditor de proteínas secretadas por vias não clássicas foi superior ao SecretomeP em todos os cenários de validação. Com relação ao PeNGaRoo, o estado da arte para este propósito, nosso preditor igualou o seu desempenho na maior parte dos cenários de validação e conseguiu acurácia melhor em alguns cenários. Dessa forma, nosso trabalho demonstrou a possibilidade da obtenção de um classificador eficaz e mais eficiente que o estado da arte, através da utilização de redes neurais artificiais e um conjunto adequado de descritores para proteínas de bactérias secretadas por vias não clássicas.

Palavras-chave: Redes neurais. Proteínas bacterianas. Secreção por via não clássica.

Abstract

We present a methodology of predicting proteins from the non-classic bacterial secretion from artificial neural networks to contribute to vaccines' elaboration and diseases' diagnosis. We compiled a list of bacterial proteins from prokaryotic organisms secreted by the currently known non-classic pathways for training. These proteins were cataloged, allowing the creation of a set of training and validation for neural network training. We carried out bibliographic research to identify probable descriptors and signaling characteristics of this type of secretion by non-classic pathway. We developed a supervised neural network using the *WEKA* software, training it to determine the best group of features for prediction. We evaluated our proposed method submitting proteins not used in the training group and comparing the developed predictor against two other predictors studied in related literature, PeNGaRoo (ZHANG et al., 2020) and SecretomeP 2.0 (BENDTSEN et al., 2005). We considered our results satisfactory, as they presented a balanced neural network accuracy of 93% in the classification performance. We outperformed SecretomeP 2.0 for all validation scenarios. In the majority, our results were similar to PeNGaRoo, but for some case studies, we outperformed it. Therefore, we demonstrated the possibility of obtaining a compelling classifier by using our selected set of descriptors.

Keywords: Neural Networks. Non-classical Secretory Pathway. Bacterial Proteins.

Lista de ilustrações

Lista de tabelas

Sumário

Capítulo 1

Introdução

A predição de proteínas secretadas por bactérias causadoras de doenças (patogênicas) é essencial, no que tange a elaboração de diagnósticos e vacinas e possui, como etapa de destaque neste processo a seleção de uma ou mais proteínas (alvos) que são secretadas pelo agente infeccioso (patógeno) (REZENDE et al., 2016).

As proteínas secretadas desempenham variadas funções em importantes processos biológicos, como: obtenção de nutrientes, mobilidade, comunicação intracelular, participação em processos de colonização no organismo hospedeiro através da liberação de toxinas e/ou enzimas degradativas, entre outras (FRONZES; CHRISTIE; WAKSMAN, 2009). Estas são encontradas em três regiões subcelulares, sendo no interior da célula (citoplasmática), aderidas na membrana plasmática, e no meio extracelular quando são exportadas completamente (OLIVEIRA et al., 2014). Na Figura 1 é possível observar uma ilustração que demostra uma proteína citoplasmática, uma de membrana e uma proteína exportada.

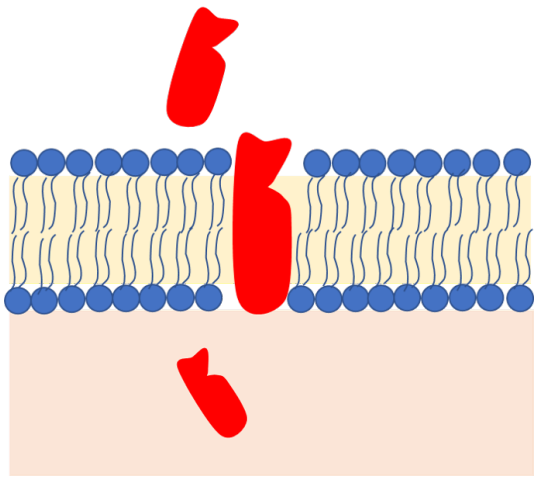

Figura 1 – Representação da posição subcelular proteica, seguindo a ordem de cima para baixo e demonstrado uma proteína secretada, de membrana e citoplasmática.

Feita pelo autor

Em todos os organismos vivos, o sistema de secreção é responsável por gerenciar a passagem de macromoléculas atravessando a membrana celular (FRONZES; CHRISTIE; WAKSMAN, 2009). Quanto aos organismos procariotos, estes criam maneiras de transportar proteínas entre locais que necessitam de sistemas de secreção especializados que, em sua maioria, são encontrados em quase todos os tipos de bactérias e são denominados como vias clássicas de secreção. No entanto, outras formas de sistemas foram identificados em um pequeno número de espécies em que apenas um número mínimo do total de todas as suas bactérias se dedicam a secretar por esta via, denominada de via de secreção não clássica (GREEN; MECSAS, 2016).

Em pesquisas sobre secreção de proteínas, as vias de secreção "Type Secretion System" (TSS) têm sido observadas em diversas bactérias (OLIVEIRA et al., 2014). As TSS podem ser divididas em sec-dependente e sec-independente (FRONZES; CHRISTIE; WAKSMAN, 2009). Dentre as vias clássicas ou sec-dependente mais conhecidas estão a Sec e Tat Pathways. Estas proteínas secretadas pela via Sec possuem aproximadamente 20 aminoácidos de comprimento em sua extremidade amino terminal, que são divididas em três regiões: um terminal amino carregado positivamente, outro terminal hidrofóbico e por fim um terminal carboxila polar. Já as proteínas secretadas pela via Tat são caracterizadas por possuírem uma sequência de aminoácido muito bem definida com um par do aminoácido arginina em sua extremidade amino terminal (GREEN; MECSAS, 2016). Estes motivos encontrados na extremidade N-terminal é denominado de peptídeo sinal e a sua presença determina a secreção da proteína por vias clássicas (NETO et al., 2012).

Essas duas vias clássicas, Sec e Tat de sinalização estão presentes na maioria das bactérias o que as tornam relativamente fáceis de serem preditas por um classificador simples que busca dentro de uma estrutura de proteína o peptídeo sinal. Entretanto, é sabido que proteínas sem esses padrões de identificação também são exportadas (sec-independente) (BENDTSEN et al., 2005). A secreção de proteínas por vias não clássicas é caracterizada por não possuir o peptídeo sinal na estrutura das proteínas exportadas, no entanto são secretadas de forma sec-independentes, este processo de secreção já observado em alguns organismos, tais como: *B. Subtilis, E. coli, Mycobacterium tuberculosis* e *Listeria monocytogenes* (BENDTSEN et al., 2005), utilizam-se de diversas vias de secreção, tornando mais complexa a identificação destas (ZHANG et al., 2020).

Contudo, é necessário conhecimento a respeito da exportação de proteína, pois além do fato delas serem utilizadas como candidatas a métodos de diagnósticos ou alvos vacinais, há ainda a possibilidade de serem utilizadas como vetores para a produção industrial de outros tipos de proteínas utilizando culturas de bactérias. Neste sentido, a ideia consiste em manipular o gene exportado incluindo a sequência de outra proteína de interesse industrial, para que este seja exportado para fora das células bacterianas e filtradas. Isso possibilita que proteínas com estrutura tridimensional formadas no interior das células bacterianas (o que não acontece com a via Sec) sejam produzidas e exportadas com

sucesso em uma cultura bacteriana (CHEN et al., 2016).

Atualmente, existem alguns métodos computacionais desenvolvidos com a finalidade de classificar proteínas secretadas por vias não clássicas, no entanto, estes métodos criados utilizam conjuntos de dados de simulação experimental, de modo geral, valendo-se de recursos básicos e treinando de maneira simples e grosseira (ZHANG et al., 2020). A exemplo disto, a remoção do peptídeo sinal das proteínas de secreção por vias clássicas, utilizando-as como conjunto positivo para o treinamento da rede neural.

Diante da importância do tema abordado e da dificuldade de se predizer se uma proteína é secretada por vias não clássicas, esta dissertação realizou um estudo que objetivou minimizar este problema abordando a relação entre as características físico-químicas dos aminoácidos na secreção de proteínas de vias não clássicas valendo-se de fundamentos da bioinformática e de técnicas de inteligência artificial, a fim de resultar na possibilidade da obtenção de um padrão para a classificação das referidas proteínas.

1.1 Motivação

Considerando-se que a predição de proteínas por vias não clássicas trata-se de terreno significantemente árido, a realização deste trabalho encontrou justificativas para elaboração de nova metodologia que viabiliza o desenvolvimento de soluções para as lacunas científicas existentes sobre tema. É sabido que, apesar da inexistência de motivos conservados (peptídeo sinal) entre proteínas secretadas por vias não clássicas, foram abordadas outras características como, número de átomos na cadeia lateral, composição proteica com predominância por certos tipos de aminoácidos (BENDTSEN et al., 2005) dentre outras. Desta forma, este trabalho de pesquisa apresenta características que possibilitaram o treinamento eficiente de um preditor que classificou proteínas secretadas por vias não clássicas, utilizando-se de redes neurais artificias, técnica de inteligência artificial que, desemboca no método proposto, possibiltando a criação de ferramentas que auxiliem na detecção e tratamento de doenças causadas por bactérias.

1.2 Objetivos

Este trabalho objetiva propor uma nova metodologia de classificação de proteínas secretadas de bactérias por vias não clássicas, evidenciando assim o conjunto de características que foi utilizado no treinamento da rede neural artificial.

1.2.1 Objetivos Específicos

O trabalho fundamenta-se na ideia de analisar características proteicas que possam indicar uma possível secreção por vias não clássicas, utilizando redes neurais artificiais

para identificar as melhores características. Sendo assim, os objetivos específicos deste trabalho consiste em:

- ❏ Elaborar um estudo bibliográfico em literaturas que abordam o tema de secreção celular em procariontes, a fim de se encontrar estruturas de proteínas que já foram identificadas como sendo secretadas por vias não clássicas;
- ❏ Estudar a bibliografia em literaturas correlacionadas para encontrar indícios de características que possam ser utilizadas como variáveis de entrada no treinamento da rede neural;
- ❏ Classificar as proteínas entre as classes de secretada ou não secretada por vias não clássicas, utilizando uma rede neural;
- ❏ Obter o melhor conjunto de "feature" efetuando a validação dos resultados obtidos por métricas estatísticas e retreinamento da rede com outros tipos de "feature" caso não se obtenha resultados satisfatórios;
- ❏ Comparar os resultados obtidos com o preditor treinado nesta pesquisa e resultados de outros preditores que compartilham o mesmo objetivo de identificar proteínas secretadas por vias não clássicas.

1.3 Organização da Dissertação

O presente trabalho foi organizado em capítulos como descrito a seguir: o capítulo 2 aborda o fundamento teórico utilizado para a criação deste trabalho descrevendo algumas técnicas sobre inteligência artificial que, por sua vez, ajudaram a compor a metodologia da pesquisa e fundamentos de técnicas com fins de avaliar o desempenho de ferramentas de classificação. O mesmo capítulo também apresenta um pequeno resumo sobre repositórios de aminoácidos e fundamentos sobre bioinformática e descreve trabalhos já desenvolvidos que compartilham de ideia correlacionada com esta dissertação. Já o capítulo 3 descreve a composição em si do método proposto, ao passo que, o capítulo 4 apresenta os experimentos e testes efetuados para avaliação de eficiência. Por derradeiro, o capítulo 5 evidencia os resultados obtidos no decorrer da pesquisa e ainda apresenta sugestões que contribuem para a complementação de resultados em trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo são abordados os conceitos teóricos que contribuíram para o desenvolvimento desta dissertação. Dentre eles estão: conceitos sobre bioinformática, aprendizagem de máquina, redes neurais, redes supervisionadas, redes neurais não supervisionadas e uma descrição sobre o banco de dados de índice de aminoácidos AAindex.

2.1 Bioinformática

A bioinformática é uma área multidisciplinar que trabalha principalmente relacionando a matemática com técnicas computacionais e estatísticas, visando contribuir com a resolução de problemas biológicos. Ademais, um dos principais objetivos da bioinformática é contribuir para a compreensão de doenças, bem como, para o desenvolvimento de tratamentos e novas drogas relacionados à essas patologias. Sabe-se que, de modo geral, pesquisas em bioinformática geram um grande volume de dados. Diante disso, o uso de técnicas baseadas em aprendizagem de máquina é vantajoso para analisar tal quantidade de dados (LORENA; CARVALHO, 2003).

No que tange ao estudo de informação biológica, a bioinformática mostra-se fundamental, pois desempenha um papel imprescindível desde a aquisição, processamento, análise, armazenamento, distribuição até a interpretação desses dados. Os resultados obtidos são integrados e auxiliam na criação de bancos de dados que proporcionam pesquisas científicas mais eficientes em áreas como biologia, agronomia, biotecnologia, dentre outras (SANTOS; ORTEGA, 2003).

As ferramentas de bioinformática tem sido utilizadas e produzidas a fim de melhorar a compreensão desse campo de estudo. Dentre as várias ferramentas produzidas ressalta-se a ferramenta *Basic Local Alignment Search Tool* (BLAST) (BORATYN et al., 2013). Essa permite comparar uma sequência de nucleotídeos ou proteína com outras amostras genômicas de domínio público identificando as regiões de similaridade entre as amostras comparadas, e ainda, utilizada na "corrida" para descobrir um medicamento para a pandemia do Covid-19 ocorrida no ano de 2020. Bem como as já citadas, a ferramenta BLAST

também foi usada em pesquisas que tinham o objetivo de identificar medicamentos disponíveis comercialmente para serem redirecionados para o tratamento do coronavírus por meio de triagem virtual baseado em estruturas (ELMEZAYEN et al., 2020).

2.2 Aprendizagem de máquina

A aprendizagem de máquina trata-se de uma vertente da inteligência artificial (IA) que tem como função desenvolver tecnologias computacionais e sistemas capazes de aprender de forma dinâmica a partir de experiências (capacidade de adquirir conhecimento durante o treinamento). Esses sistemas têm características de tomar decisões próprias de acordo com experiências na execução em tempos anteriores (MONARD; BARANAUSKAS, 2003). As principais abordagens utilizadas na inteligência artificial (AI) são: aprendizagem de máquina supervisionada, não supervisionada e aprendizagem por reforço.

A primeira se caracteriza pelo fato de um sistema receber uma tarefa com sua resposta já preestabelecida. Ou seja, pode-se dizer que ele recebe um conjunto de atributos com valores de entradas e saídas (rótulos) (BAILEY; HARRIS, 1985). Esse tipo de aprendizado supervisionado tem como característica induzir o resultado do processamento (BATISTA et al., 2003). Já na segunda abordagem, chamada de não supervisionada, o sistema, recebe apenas os atributos de entrada (BAILEY; HARRIS, 1985) e tem por objetivo construir um modelo que procura padrões em comum entre as amostras disponibilizadas, formando assim grupos que possuem características similares (BATISTA et al., 2003). Esta técnica é conhecida como algoritmos de clusterização podendo ser citado como exemplo o algoritmo *k-Means* (HONDA; FACURE; YAOHAO, 2017).

A aprendizagem por reforço (AR) é um modelo computacional de aprendizado de máquina onde o algoritmo tem a função de melhorar seu desempenho baseado em um reforço (recompensas ou punições) que recebe ao fim da interação com um ambiente desconhecido (RIBEIRO et al. 2006). Assim sendo, a aprendizagem por reforço utiliza um crítico externo ao ambiente que avalia os resultados do algoritmo, mas não indica explicitamente o resultado correto (GUELPELI; RIBEIRO; OMAR, 2003).

Tendo em vista, todas as abordagens apresentadas, a arquitetura supervisionada é escolhida para compor o desenvolvimento do trabalho, para assim avaliar a sua eficiência juntamente com os conjunto de entrada e saída na classificação de possíveis proteínas que serão secretadas por vias não clássicas. E desta forma definir se esta arquitetura é eficiente para este problema apresentado.

2.3 Redes neurais artificiais

A inteligência foi a principal característica que possibilitou a adaptação e a sobrevivência da espécie humana na natureza e, por consequinte, uma das áreas que desperta

a atenção de muitos pesquisadores é a simulação das capacidades cognitivas do ser humano. O elemento-chave para a inteligência do ser humano é o cérebro, responsável por uma rede de células neuronais que geram e transmitem impulsos elétricos. Tais impulsos são incumbidos de manter a troca de informações biológicas que sustentam a capacidade cerebral. Embasados neste sistema, os pesquisadores se basearam na estrutura do cérebro para criar um ambiente técnico chamado de redes neurais computacionais (SEGATTO; COURY, 2006).

A pesquisa sobre redes neurais artificiais datam do início do ano de 1940 com o trabalho de (MCCULLOCH; PITTS, 1943). Entretanto, somente em 1950 e 1960 que pesquisadores como (ROSENBLATT, 1958) se aprofundaram no tema, sendo este último o primeiro a propor um modelo inovador de redes neurais chamado de *perceptron*. Contudo, dificuldades metodológicas e tecnológicas levaram esse modelo ao esquecimento. Somente na década de 80 os estudos desse modelo retomaram graças aos avanços nos modelos metodológicos e recursos computacionais (FERNEDA, 2006). A Figura 2 ilustra o modelo matemático de um *perceptron* simples.

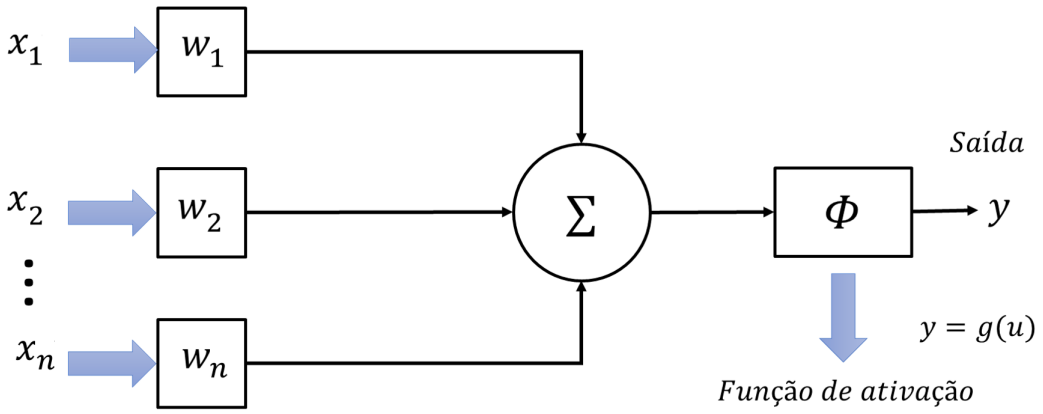

Figura 2 – Modelo *perceptron* simples

Figura adaptada de: (FERNEDA, 2006)

o modelo ilustrado na Figura 2 é composto por três elementos básicos:

Conjunto das n conexões de entrada (x_1,x_2,...,x_n), e pesos (w_1,w_2,...,w_n);

O somador (Σ) cuja função é somar a interação dos conjuntos w e x;

A função de ativação (ϕ), representada pela função g que tem como parâmetro u, resultado de (Σ);

y que representa a saída do neurônio (FERNEDA, 2006);

.

O conjunto dos pesos possui um papel fundamental no modelo, pois é ele que define o comportamento das conexões entre os neurônios, enquanto o conjunto das conexões de en-

trada (x) é multiplicado pelo conjunto de pesos (w) somado pelo somador posteriormente (Σ), e o valor resultante, enviado para a função de ativação.

Esta combinação de vários neurônios forma uma rede neural artificial que visa simular uma rede neural biológica onde os neurônios se unem por meio de conexões sinápticas e processam informações. A Figura 3 ilustra uma rede neural computacional que se assemelha a um grafo onde os nós são os neurônios e as ligações têm a função de efetuar as sinapses (FERNEDA, 2006).

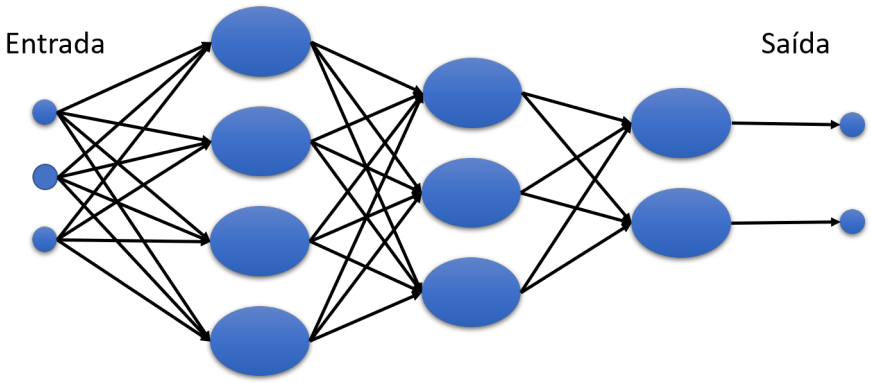

Figura 3 – Representação de uma rede neural artificial

Figura adaptada de: (FERNEDA, 2006)

2.4 Aprendizagem supervisionada

Uma rede supervisionada (RS) é baseada num conjunto de exemplos de estímulo resposta, ou em algum outro tipo de informação, que represente o comportamento que deve ser apresentado pela rede neural (HAYKIN, 2007). Deve-se fornecer à rede tanto os parâmetros de entrada quanto o valor desejado de saída para cada amostra de entrada, tornado possível a comparação do valor obtido na saída da RS com a saída desejada, obtendo assim o erro referente a resposta atual. Com este erro calculado é possível treinar a rede reajustando os pesos sinápticos de forma a minimizar o erro (FLECK et al., 2016).

Um exemplo de rede neural cujo treinamento se insere no contexto supervisionado é o Perceptron Multicamadas ou MLP (*Multi Layer Perceptron*). Essa rede é constituída de neurônios que podem ser combinados entre si em uma estrutura em camadas em que, cada camada possui número diferente de neurônios, e ainda pode conter várias camadas, tornando-a profunda e capaz de aprender relações cada vez mais complexas. O MLP é caracterizado pelo fato de conter uma camada de entrada, uma ou mais camadas intermediárias ou camadas ocultas e uma camada de saída (HAYKIN, 2007).

O treinamento de uma rede é uma etapa crucial, pois determina se ela obterá sucesso ou fracasso considerando vários fatores como o algoritmo de treinamento e o número de

épocas (iterações para obter o resultado ótimo) (GUIMARÃES et al., 2008). O fluxo de treinamento da rede MLP é descrito pelo vetor de valores de entrada que passam pela camada inicial, cujos valores de saída são ligados às entradas da camada seguinte, e assim por diante, até a rede fornecer como resultado os valores de saída da última camada. A saída da rede é comparada com um valor já preestabelecido (saída desejada) e o resultado da diferença entre ambos fornece o valor do erro. Este valor do erro é retro propagado entre as camadas anteriores, inicia-se pela camada de saída e passa-se por todas as camadas até a camada de entrada, onde com este erro obtido é possível recalcular os pesos sinápticos (MANZAN et al., 2016). Essa é a descrição de um dos algoritmos mais utilizados para efetuar o treinamento de uma rede MLP, o chamado de *backpropagation*, que em português significa retropropagação e consiste na seguinte ideia: com base no cálculo do erro ocorrido na camada de saída da rede neural, recalcular o valor dos pesos da última camada de neurônios e assim proceder para as camadas anteriores, ou seja, atualizar todos os pesos das camadas a partir da última até atingir a camada de entrada. Em outras palavras, calcula-se o erro entre o que a rede encontrou com o valor desejado de fato. Então, recalcula-se o valor de todos os pesos, começando da última camada e indo até a primeira, sempre tendo em vista diminuir esse erro. Salienta-se que, neste algoritmo é definida a taxa de aprendizagem da rede, o número de neurônios na camada oculta e o critério de parada do treinamento (MANZAN et al., 2016).

Este tipo de rede neural possui a capacidade de resolver diversos tipos de problemas característicos como aproximações, previsões, classificações, entre outros (BRAGA; FERREIRA; LUDERMIR, 2007). Deste modo, no campo da bioinformática eles podem ser encontrados em problemas que envolvem classificação de proteínas, no reconhecimento de sinais, identificação de assinaturas, identificação de repetidores de uma rede de baixa complexidade, similaridades entre sequências, análise de cromatogramas, predição de estruturas secundárias de proteínas (COELHO, 2016).

2.5 Aprendizagem não supervisionada e Aprendizagem competitiva

A rede neural não supervisionada se caracteriza pelo fato de suas amostras de treinamento não conterem os atributos que representam a saída desejada. Desse modo, a rede terá o desafio de tentar classificar seus dados sem a ajuda de um agente externo. Neste contexto, pode-se citar a rede neural por aprendizado competitivo. Ela é caracterizada pelo fato dos neurônios de uma mesma camada competirem entre si em busca de obter um único neurônio vencedor, sendo o neurônio que tiver o maior grau de atividade (BARRETO, 1998). Assim, em cada neurônio seu nível de ativação multiplica o valor de entrada pelos seus respectivos pesos sinápticos para poder identificar o neurônio vencedor que será aquele cujo valor do nível sináptico calculado é o maior. No aprendizado

hebbiano, vários neurônios da camada de saída podem estar simultaneamente ativos, mas na aprendizagem competitiva apenas um neurônio fica ativo de cada vez. Dito isto, os principais fundamentos para a existência de uma rede competitiva são:

- ❏ A existência de um conjunto de neurônios idênticos ligados por valores de conexão sináptica com valores distribuídos aleatoriamente.
- ❏ A definição de um valor máximo para a ativação do neurônio.
- ❏ A criação de um mecanismo que permita os neurônios entrarem em competição na busca de apenas um permanecer excitado (JORGE, 2013).

O objetivo do aprendizado competitivo é fazer com que os neurônios se especializem de forma a não supervisionar os seus estímulos, pois nenhuma informação sobre a classe que está sendo estimulada é passada no processo de ajuste de peso sináptico (COSTA et al., 1999).

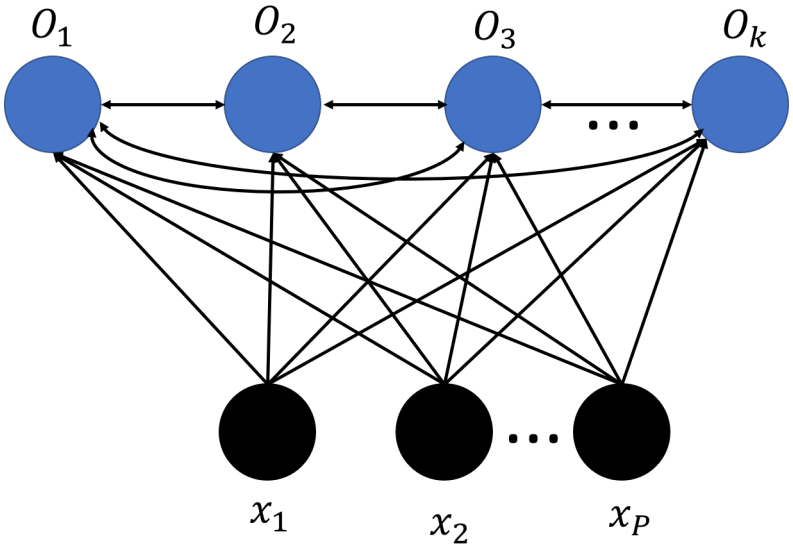

Figura 4 – Modelo rede neural competitiva.

Figura adaptada de: (COSTA et al., 1999).

A Figura 4 ilustra um modelo simples de rede neural competitiva composta pelo vetor de entrada da rede X normalizado ($||X|| = 1$). Os neurônios da camada de saída são representados pelo vetor O_i sendo $i = 1, 2, 3,..., k$, eles têm a função de calcular o valor de ativação. O vencedor será aquele que apresentar o máximo valor de ativação, isso significa que no final do processamento apenas um ficará ativo (COSTA et al., 1999). O modelo de rede auto organizável de Kohonen é um modelo clássico de rede competitiva (KITANI, 2013).

2.6 AAindex

AAindex (Amino Acid index database) é um banco de dados de índices de aminoácidos. Cada índice é composto por vinte valores numéricos que representam as propriedades físicas, químicas e bioquímicas dos aminoácidos. O repositório contém também, índices compostos por pares de aminoácidos que formam matrizes que podem ser simétricas e compostas de 400 valores numéricos (20x20).

O AAindex foi estruturado a partir de um esforço de pesquisas experimentais e teóricas para se organizar os diferentes tipos de propriedades de aminoácidos e caracterizá-los em índices numéricos. A pesquisa, iniciada em 1988, coletou 222 amostras de índices de aminoácidos que foram identificadas em quatro grupos principais: (i) *α-helix and turn propensities*, (ii) *β-strand propensity*, (iii) *hydrophobicity* e (iv) *physicochemical properties*.

O repositório AAindex pode ser acessado através do serviço japonês GenomeNet no link (AAINDEX, 2020) [1]. Ele é dividido em três seções: AAindex1 que armazena os índices de aminoácidos, AAindex2 que guarda as matrizes de mutações e a seção AAindex3 que armazena as matrizes de proteínas de contato de aminoácidos. Na publicação desta ferramenta a base AAindex1 continha 434 índices de aminoácidos mapeados (KAWASHIMA; KANEHISA, 2000). Na presente data da construção deste trabalho no ano de 2020 a base do AAindex contém 566 índices de aminoácidos.

2.7 Métodos de validação

Neste sub item é descrito as técnicas Curva de Roc e Validação cruzada onde elas contribuíram para medir a eficiência da rede neural artificial treinada.

2.7.1 Curva de ROC

O gráfico de ROC (do inglês Receiver Operating Characteristic) tem a vantagem de fornecer uma representação da sensibilidade e especificidade dos valores limites que são invariáveis por meio de transformações monótonas. Ele permite comparar dois ou mais classificadores analisando suas capacidades discriminativas (BENAVIDES, 2017).

Fundamentalmente, este gráfico trata de um método de avaliação e organização de resultados de diagnósticos, predições e outras aplicações em diferentes áreas, como por exemplo, em telecomunicações na detecção da qualidade de um sinal de uma transmissão de um canal com ruídos; na psicologia com a avaliação da capacidade de pacientes em distinguir um estímulo; na medicina com a melhoria da avaliação da qualidade de um teste clínico; na economia com a avaliação da desigualdade de renda e; na metereologia com a avaliação de eventos raros climáticos (PRATI; BATISTA; MONARD, 2008).

[1] Disponível em: <http://www.genome.ad.jp/aaindex> Acessado em 20 dezembro 2020

Pode-se exemplificar a aplicação de ROC utilizando um algoritmo de aprendizado supervisionado que possui um conjunto de atributos distintos para o treinamento rotulados em positivo ou negativo, assumindo que o algoritmo é um classificador binário. Considerando uma análise de dois possíveis resultados de um exame clínico, os cenários possíveis são: um indivíduo doente que recebe um resultado positivo na análise é um verdadeiro positivo (VP); um indivíduo saudável que recebe um valor negativo é um verdadeiro negativo (VN); já um indivíduo que recebe um resultado positivo, mas está saudável é considerado um falso positivo (FP); por último, um indivíduo que recebe um resultado negativo na análise, mas que está doente, é considerado um falso negativo (FN). A curva de ROC permite visualmente fazer a distinção entre sensibilidade (S) eixo das ordenadas e especificidade (E) eixo das abcissas, onde esta curva é obtida através do cálculo da sensibilidade e especificidade (CRISTIANO, 2017). De modo breve, pode-se ressaltar que:

- A Sensibilidade demonstrada na equação 1 também pode ser considerada como a taxa dos positivos verdadeiros VP. Ela representa a taxa da quantidade de indivíduos que foram classificados corretamente.

- A Especificidade demonstrada na equação 2 representa a taxa dos verdadeiros negativos VN. Ela representa a taxa dos indivíduos que foram classificados como negativo corretamente para uma determinada classificação (CARMINATI et al., 2003) (pfs.io,2017). Fórmulas retiradas de (CRISTIANO, 2017).

$$S = \frac{VP}{VP - FN} \tag{1}$$

$$E = \frac{VN}{VN - FP} \tag{2}$$

Com isto, podemos identificar um preditor perfeito quando ele é descrito 100% sensível, ou seja, quando todos os indivíduos positivos são classificados corretamente como positivos (considerando a analogia de um exame onde um indivíduo saudável é classificado como saudável). Sendo assim, quando ele é 100% específico todos os indivíduos negativos são classificados como negativos (indivíduos doentes são identificados como doentes) (CARMINATI et al., 2003)

2.7.2 Validação cruzada

A grande dificuldade do treinamento das rede neurais é atingir o melhor ponto de parada, tendo assim uma boa generalização, ou seja, a capacidade de predizer corretamente onde o erro do treinamento inicia com um valor alto e tende a atingir o valor mínimo local (GUIMARÃES et al., 2008).

Desta forma a validação cruzada consiste em um modelo estatístico com o intuito de validar os resultados da rede utilizando um conjunto de dados diferente para o treinamento e validação. Esta técnica acompanha a evolução do aprendizado da rede observando a curva dos subconjuntos de dados de treinamento e validação. Assim, o fluxo de treinamento é interrompido quando a curva de validação é menor que o erro mínimo e antes que a curva retome seu crescimento (GUIMARÃES et al., 2008).

O fundamento da validação cruzada se concentra em particionamento do conjunto de dados divididos em N subconjuntos (conjunto/N onde N é a quantidade de Folds) mutuamente exclusivos e, posteriormente alguns destes conjuntos são utilizados para o treinamento do modelo, enquanto os subconjuntos restantes são utilizados em dados de validação ou testes no modelo treinado, processo repetido N vezes (SINGH; PANDA, 2011).

2.8 Trabalhos relacionados

Nesta seção, aborda-se trabalhos que possuem teorias semelhantes e estão correlacionadas com o tema dessa dissertação.

(ZHANG et al., 2020) desenvolveu um trabalho cujo objetivo foi criar um preditor que possa identificar se uma determinada proteína é secretada por vias não clássicas. Inicialmente, os autores buscam a partir de outros trabalhos relacionados (WANG et al., 2016),(BENDTSEN et al., 2005) uma coleção de proteínas para treinamento e validação de um preditor. Neste caso são utilizadas proteínas secretadas por vias não-clássicas de bactérias gram-positivas. São extraídas 253 sequências de proteínas secretadas por vias não-clássicas; já para o conjunto de treinamento negativo utilizam-se 1084 proteínas. Essas amostras são enviadas para um filtro a fim de remover qualquer redundância e também são divididas em dois conjuntos de dados, sendo 90% das amostras positivas para treinamento e os 10% restantes para validação.

Os autores utilizaram-se de diversos grupos de características para treinamento, o primeiro deles ou grupo 1 é composto de pseudo-aminoácidos (PAAC) sendo seus descritores (hidrofobicidade, hidrofilicidade e a cadeia lateral de massas dos 20 aminoácidos), ordem quase-sequência (QSO), que mede a ocorrência de aminoácidos com base na matriz de distância físico-química de Schneider-Wrede e a matriz de distância química de Grantham. Já o grupo 2 foi uma evolução das características PAAC que, utilizou-se do software POSSUM para gerar outras características como TPC, Pse-PSSM e AATP. Finalmente, no grupo 3 concentra as características físico-químicas, sendo elas conjoint Triad (CTriad) e transição entre CTD (CTDT).

Para treinamento e otimização foi utilizado o algoritmo LightGBM com aumento de gradiente (LIGHTGBM, 2020) [2]. Com intuito de avaliar rigorosamente o desempenho

[2] Disponivel em: <https://github.com/Microsoft/LightGBM> Acessado em 20 dezembro 2020

da previsão dos modelos propostos aplicou-se três tipos de testes de validações: k-fold, leave-one-out e independent test. Para avaliar o desempenho de forma abrangente e quantitativa foram utilizadas as medidas de Sensibilidade (SN), de Especificidade (SP), de Precisão (ACC), F-value e o Coeficiente de correlação de Matthew (MCC) que são os métodos mais utilizados na bioinformática (ZHANG et al., 2020). Estas medidas podem ser definidas pelas seguintes fórmulas:

$$SN = \frac{TP}{TP + FN} \tag{3}$$

$$SP = \frac{TN}{TN + FP} \tag{4}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$F - Value = 2 \times \frac{TP}{2TP + FP + FN} \tag{6}$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{7}$$

Os valores TP, TN, FP e FN representam os números de verdadeiros positivo, verdadeiros negativo, falsos positivo e falsos negativo, respectivamente.

Na avaliação dos resultados os autores concluíram que os modelos de treinamentos que utilizaram o PSSM (ou seja, TPC, Pse-PSSM e AATP) tiveram uma melhor solução se comparados com os modelos de atualização de propriedades físico-químicas baseadas em sequência. Este preditor foi disponibilizado através de um web service que se encontra disponível para consultas públicas (PENGAROO, 2020) [3]. Desta forma, pode-se enviar amostras e receber a classificação das mesmas online. Para fins de otimização a saída da resposta da submissão pode assumir os seguintes rótulos: “Exp” se a proteína já for conhecida como secreção não clássica, e se ela for submetida para predição ela é rotulada como “Pred” (ZHANG et al., 2020).

A pesquisa conduzida por Jannick D Bendtsen, Lars Kiemer, Anders Fausboll, e Soren Brunak (BENDTSEN et al., 2005) descreve um trabalho com uma visão genérica de secreção de proteínas por vias não clássicas em bactérias. Os autores apresentam um método de predição independente do peptídeo sinal e, no decorrer do desenvolvimento do trabalho constroem uma tabela com uma lista de proteínas encontradas de forma extra celular que não tenham um peptídeo sinal. A metodologia empregada efetua uma pesquisa bibliográfica para reunir um conjunto de proteínas já conhecidas por serem secretadas por vias não clássicas e aplicar métodos de identificação de padrão que pudessem identificar a secreção por vias não clássicas.

[3] Disponivel em: <https://pengaroo.erc.monash.edu/> Acessado em 20 dezembro 2020

Segundo o trabalho, há uma grande dificuldade de realizar uma pesquisa deste tipo, pois pouco se sabe sobre os padrões de secreção por vias não clássicas dificultando o processo de classificação correta por métodos preditivos. O trabalho também descreve o desenvolvimento de uma rede neural com base em características, tais como, a modificação pós tradução, as estruturas secundárias, os pontos isoelétricos e o comprimento de sequência e a utilização de proteínas de secreção por vias clássicas para compor o conjunto de treinamento. Logo, a ideia seria buscar evidências analisando suas características desenvolvendo dois preditores: um para classificação de proteínas de bactérias Gram-positivas e outro para as proteínas de bactérias Gram-negativas.

Essa pesquisa demonstrou que é possível identificar proteínas não clássicas em mamíferos utilizando como propriedade de avaliação suas características biológicas e químicas e as comparando com características de proteínas classicamente secretadas já conhecidas. Devido ao fato de não existir uma base de proteínas secretadas por vias não clássicas conhecidas foi efetuada a remoção do peptídeo sinal das proteínas classicamente secretadas para efetuar o treinamento da rede. Em seguida foi apresentada uma lista compilada de proteínas secretadas por vias não clássicas através de uma exaustiva pesquisa bibliográfica, assim como um novo método de predição de proteínas secretadas por vias não clássicas em bactérias Gram-positivas e Gram-negativas (BENDTSEN et al., 2005). Neste trabalho foi desenvolvido uma rede neural utilizando as características dos aminoácidos como atributo de treinamento da rede, das quais estão descritas na Figura 5.

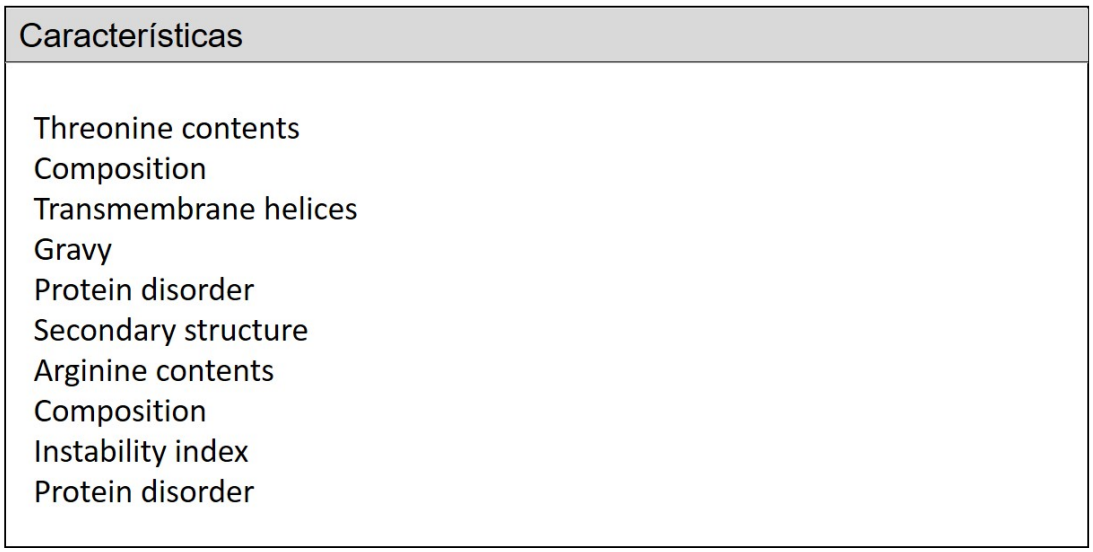

Figura 5 – Características utilizadas para o treinamento do preditor.

(BENDTSEN et al., 2005).

Para este treinamento obteve-se amostras de proteínas Gram-positivas e Gram-negativas, e para tanto foram filtrados respectivamente os conjuntos 152 e 350 para treinamento positivo e 140 e 334 para treinamento negativo. Uma validação cruzada tripla foi utilizada

para garantir que os resultados de treinamento e o coeficiente de correlação determinados obtivessem o melhor desempenho. O método atribui um valor de saída de 0 e 1 e se a pontuação da saída da rede for superior a 0,5 é considerado provável a secreção (BENDTSEN et al., 2005).

2.9 Conclusão

Na literatura recente verifica-se progressos na compreensão deste fenômeno (KANG; ZHANG, 2020) e um avanço para se obter um preditor confiável que determine se a proteína é secretada por vias não clássicas utilizando-se de características físico-químicas, assim como a composição dos aminoácidos (ZHANG et al., 2020). Contudo, apesar dos esforços um preditor ideal ainda não foi obtido. À medida em que novas pesquisas surgem melhorias são feitas e agregadas às teorias já existentes e publicadas, incluindo a linha de pesquisa presente, que encontra-se em aberto com uma vasta possibilidade de teorias a serem abordadas. É neste contexto que o presente trabalho se posiciona para evidenciar um conjunto de características e, predizer com eficiência a classificação de proteínas por vias não clássicas, utilizando-se de técnicas de inteligência artificial para a descrição de características que sinalizem se uma proteína pode ser secretada por vias não clássicas.

Capítulo 3

Metodologia

A proposta desta pesquisa é utilizar técnicas de inteligência artificial que permitam encontrar um conjunto de características para predizer Proteínas Secretadas por Vias não Clássicas (PSVnC's). A metodologia é dividida em quatro etapas, (i) busca de proteínas na literatura, (ii) busca de características físico-quimicas sinalizadoras de secreção de PSVnC's na literatura, (iii) criação, treinamento e testes de uma rede neural utilizando a biblioteca de software WEKA (FRANK et al., 2004), (iv) avaliação e comparação dos resultados.

3.1 Revisão literária

A primeira etapa do trabalho consistiu em efetuar uma revisão da literatura sobre PSVnC's, a qual teve seu foco a leitura de trabalhos e ferramentas relacionadas ao tema de secreção de proteínas (WANG et al., 2016). Uma vez identificadas PSVnC's, essas foram catalogadas permitindo a criação de um conjunto de treinamento e validação de uma rede neural. Salienta-se que, uma parcela das PSVnC's foi obtida dos dados públicos referentes aos programas feitos com o mesmo propósito sendo SecretomeP-2.0 (BENDTSEN et al., 2005) e PeNGaRoo (ZHANG et al., 2020), pois estas ferramentas disponibilizam dados de amostras de proteínas. O controle negativo foi composto de proteínas de secreção clássica, membranas obtidas do repositório UNIPROT em formato FASTA.

Utilizou-se um programa desenvolvido para auxiliar neste trabalho denominado *valifasta* que, remove caracteres de pontuação, quebras de linha e outros prováveis caracteres ocasionalmente inseridos em sequências de proteínas e ainda, garante que uma proteína tenha uma chave de identificação única dentre as contidas em um mesmo arquivo denominado *multifasta*. Tais medidas visam minimizar prováveis erros de processamento pela utilização de ferramentas de bioinformática que não aceitem outro formato senão as vinte letras dos aminoácidos e uma identificação simples para cada proteína. Na Figura 6 há uma representação de como uma proteína em formato FASTA é convertida em formato de linha após utilização do *valifasta*.

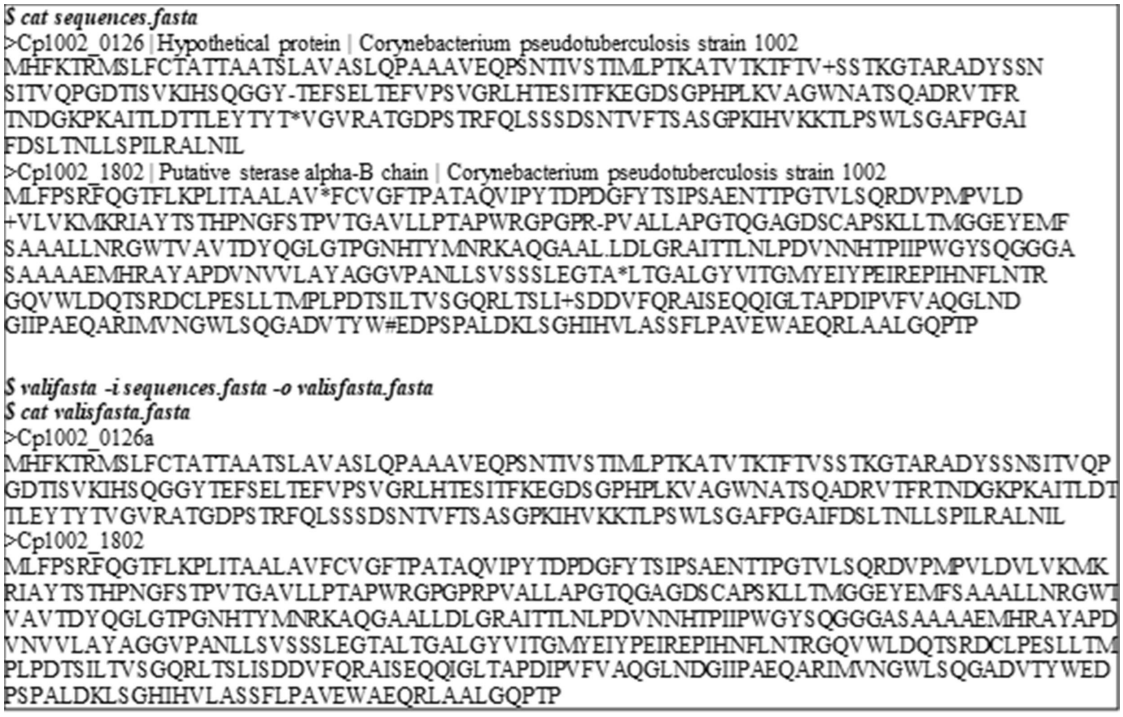

Figura 6 – Representação das proteínas em formato fasta antes e após conversão com o valifasta.

Elaborada pelo autor.

3.2 Busca por descritores

A segunda etapa consistiu em buscar prováveis descritores/características sinalizadoras de secreção por vias não clássicas em proteínas de células procarióticas e ainda, em extrair informações de artigos como (HUNG et al., 2010) indicando possíveis características químicas ou físicas sinalizadoras de secreção, exemplificando: a proporção de uma proteína com estrutura tridimensional alfa-hélice, características físico-químicas diversas como hidrofobicidade, hidrofilicidade, quantidade de aminoácidos carregados negativamente ou positivamente e tipos de metais presentes em sítios catalíticos. Neste mesmo momento utilizou-se do conhecimento estabelecido que PSVnC's são secretadas pela membrana plasmática levando em consideração o nível de afinidade entre membranas e proteínas levando em conta que a membrana plasmática de procariotos é constituída principalmente de fosfolipídios com regiões hidrofílicas e hidrofóbicas. Assim, a proporção de aminoácidos com essas características será considerada na seleção de alvos para o conjunto de treinamento. Pelo fato de ser membrana predominantemente composta de aminoácidos hidrofóbicos, esta hidrofobicidade permite que as moléculas de água do interior e do exterior de uma célula bacteriana contribuam para aglomerar e comprimir esses aminoácidos em uma barreira que constitui o invólucro de uma célula bacteriana. Desse modo, não é de se esperar que PSVnC's tenham predominância de aminoácidos

hidrofóbicos. Desta forma, a hidrofilia e a hidrofobia são exemplos de descritores a serem utilizados no processo de treinamento da rede neural proposta no presente trabalho.

Ainda quanto a geração de descritores foram utilizados quatro grupos de classificação de aminoácidos conforme a Figura 7, em que , aminoácidos de característica básica são representados pela cor verde, não polar (hidrofóbicos) são representados pela cor azul, polar (hidrofílicas) estão em roxo e ácidas são representados pela cor alaranjada. Além disso, foi criada outras características para o treinamento da rede levando-se em consideração a quantidade de aminoácidos com as mesmas características químicas (polar, não polar, básica, ácida).

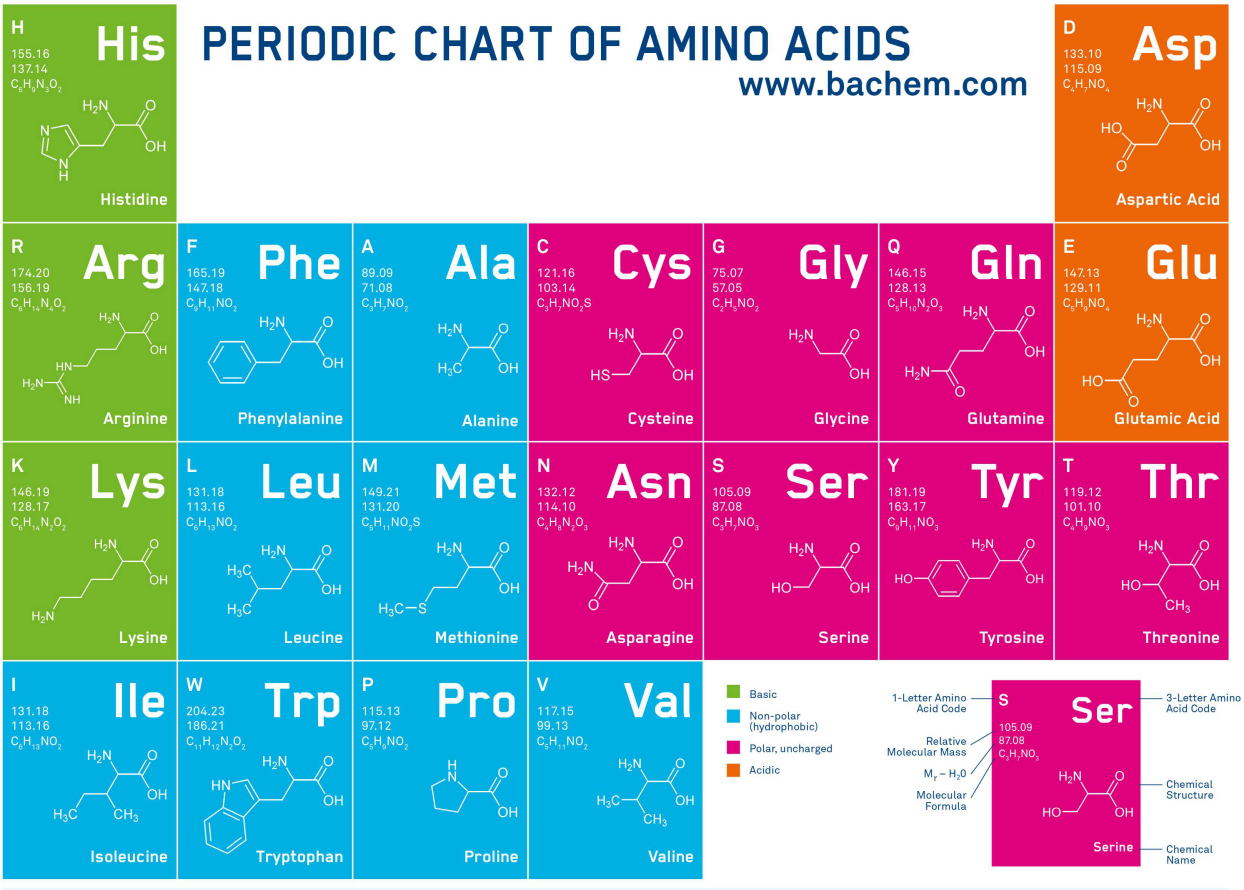

Figura 7 – Tabela química dos aminoácidos.

(MERGLER, 2020)

É sabido que, por meio de experimentos efetuados neste trabalho que apenas esses quatro descritores não são suficientes para classificar sequer os quatro locais celulares mais comuns (citoplasma, secreção clássica, membrana e exposto na membrana). Por esse motivo, buscou-se novos descritores no repositório "AAindex1" (KAWASHIMA; KA-

NEHISA, 2000). Esse repositório (AAINDEX, 2020)[1] contém, até a data da escrita desse projeto, várias centenas de índices de propensão para os 20 aminoácidos mais comuns que, consideram médias numéricas para características químicas, físicas e estruturais de aminoácidos em um conjunto de proteínas significativamente representativo dos organismos conhecidos. A seguir, a Figura 8 apresenta uma captura de tela de um índice de propensão de aminoácidos retirada do AAindex utilizados neste trabalho.

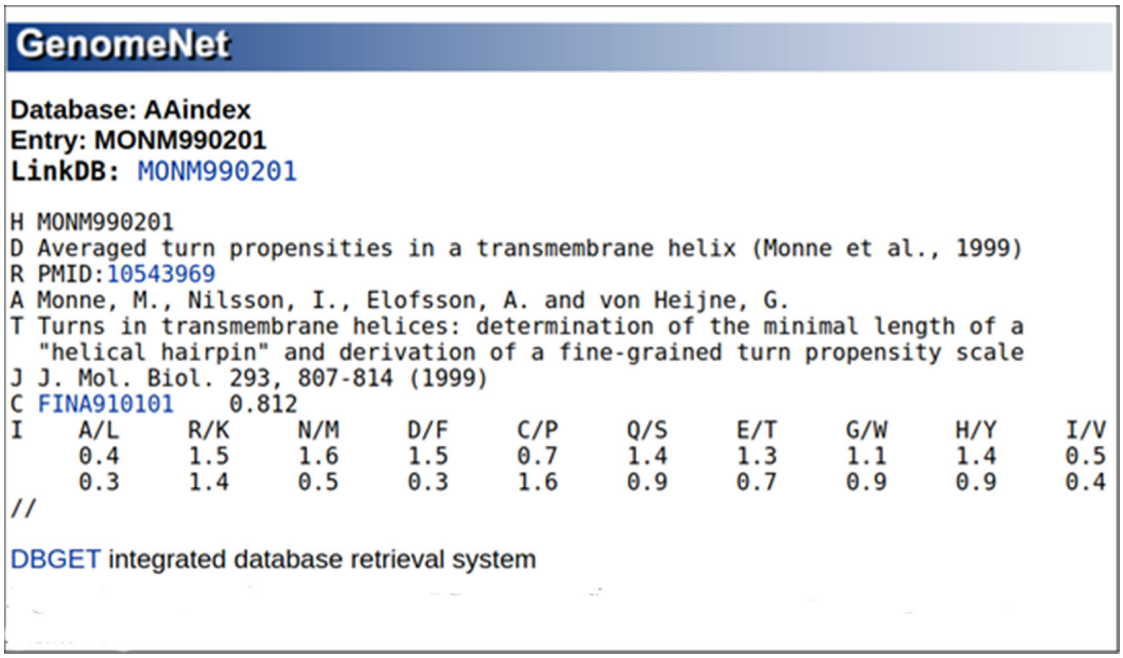

Figura 8 – Exemplo de índice de propensão de aminoácidos disponíveis no repositório AAindex.

3.3 Treinamento e testes da Rede Neural

A terceira etapa utilizou-se do software WEKA, onde foi criada uma rede neural supervisionada que treinou um modelo cujo objetivo foi encontrar as melhores características que identifiquem se uma proteína é secretada por vias não clássicas. Essa etapa utilizou o conceito de Validação Cruzada (VC), a qual divide o conjunto de treinamento em N subgrupos, treina um modelo de aprendizado em N-1 grupos e testa em um grupo. O processo se repete até que os N subgrupos tenham assumido o papel do grupo de testes. O desempenho do modelo de aprendizado final é uma média dos resultados dessas N execuções. Assim, nessa etapa acontece o processamento das características obtidas na etapa anterior para o treinamento da rede neural com o software WEKA. Tal processamento

[1] Disponivel em: <https://www.genome.jp/aaindex/> Acessado em 20 dezembro 2020

consiste em calcular a frequência de aminoácidos das cadeias de proteínas de acordo com os descritores eleitos para alimentar a rede neural. As frequências dos 20 aminoácidos e os valores atribuídos de acordo com o AAindex servem como pesos que, permitem criar vetores numéricos para cada proteína de acordo com os descritores selecionados. A Tabela 1 a seguir, enumera alguns dos pesos que representam as características de treinamento selecionados para esta pesquisa (o conjunto completo se encontra nos anexos A.3).

Tabela 1 – Índices de propensão de aminoácidos utilizados para gerar descritores de proteínas na busca por classificar proteínas secretadas por vias não clássicas.

Reference	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
BASIC	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ACID	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
POLAR	0.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0
NON-POLAR	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	1.0
CHAM830103	0.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	2.0	1.0	1.0	1.0	1.0	0.0	1.0	2.0	1.0	1.0	2.0
CHAM830104	0.0	1.0	1.0	1.0	0.0	1.0	1.0	0.0	1.0	2.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0

Para formatar os dados em um modelo de treinamento do software WEKA foi desenvolvido um programa chamado *features* (Linguagem Lisp), disponível no Anexo I. O programa *features* processa um arquivo multifasta gerando um arquivo CSV (*Comma-separated values*) contabilizando os valores numéricos de todos os descritores selecionados (Figura 9). Os descritores selecionados da Tabela 1, e seus respectivos pesos para os 20 aminoácidos, são armazenados em um arquivo-texto de modo a evitar a recompilação do programa *features* a cada descritor alterado.

```
ALRKNMDFCPQSETGWHYIVBASICACIDPOLARNON.
POLARPHAM830103CHAM830104CHAM830105FAUJ880111MONM990201MONM990101011INIBASICINIACIDINPOLARININON.
POLARINICHAM830103INICHAM830104INICHAM830105INIFAUJ880111NIMONM990201INIMONM99010IENDBASICENDACIDENDPOL
ARENDNON-
POLARENDCHAM830103ENDCHAM830104ENDENDCHAM3010SENDFAUJ880111ENDMONM990201ENDMONM990101MIDBASICMIDACID
MIDPOLARMIDNIN,POLARMIDCHAM830103MIDCHAM830104MIDCHAM830105MIDFAUJ880111MIDMONM990201MIDMONM990101
Cp1002_0126a208791671451118123111328342415 25.016.094.093.0241.0114.059.025.0199.00002202.999983.01.015.021.010.016.010.
03.030.030.46.02.023.029.059.034.01256.051.753.10000211.010.029.026.078.043.024.511.073.377.59999
Cp1002_1802451617193181433872045610123727357132929.033.0148.0205.0379.0221.0106.529.0368.1389.899932.01.012.025.038.020.011
.02.030.832.45.07.017.031.051.034.020.55.054.759.110.09.053.067.0121.076.035.010.0120.50001128.0
```

Figura 9 – Arquivo CSV resultado do processamento do arquivo fasta pelo programa *features*. Esse formato é genérico. Um script bash ainda precisa converter o CSV para o formato ARFF do programa *WEKA*.

Esse arquivo CSV de exemplo possui três linhas: uma de cabeçalho e duas de dados. A linha de cabeçalho lista os descritores na cor verde; o rótulo de cada proteína está em azul e os números referentes a cada descritor estão na cor preta. Para cada descritor da Tabela 1 são derivados outros três que analisam as porções de início, meio e fim de cada

proteína. O motivo dessa derivação de descritores reside no fato que, de acordo com a localização celular, proteínas podem ter padrões de aminoácidos com características físico-químicas distintas nessas porções protéicas. A título de exemplo, as proteínas secretadas por vias clássicas geralmente possuem padrões hidrofóbicos apenas em sua porção inicial. O arquivo CSV é formatado para um ARFF, padrão de entrada de dados do software WEKA, por meio de um script bash do SO Linux (Figura 10).

```
features valifastafasta> weka.arff
head -n 1 weka.arff > weka.attributes
end=`head -n 1 weka.arff | wc -w`; sed -i "s/\t$//g"  weka.arff; echo "$end features"
sed -i '1d' weka.arff
sed -i "s/\([a-zA-Z0-9]\+\)/@attribute \1 numeric#/g" weka.attributes
n '#' '\n' < weka.attributes > weka.attributes2
n -d '\t' < weka.attributes2 > weka.attributes
/addlocal local weka.arff #incere rótulos das instâncias de dados
cut -f 2- weka.arff > weka.arff2
sed -i "s/\t/,/g" weka.arff2
sed -i "s/$/,\?/g" weka.arff2
echo '@relation localsubcellular' > localsubcellular.arff
cat weka.attributes  >> localsubcellular.arff
echo '@attribute class {POSITIVE, NEGATIVE}' >> localsubcellular.arff
echo '@data' >> localsubcellular.arff
cat weka.arff2  >> localsubcellular.arff
```

Figura 10 – Programa em bash para converter o CSV genérico para ARFF utilizado pelo *WEKA*.

Dessa forma, converte-se uma proteína com representação de aminoácidos em letras para uma cadeia de números, padrão WEKA de entrada de dados, com as características sinalizadoras de secreção não clássica. Desta feita, a representação do rótulo de cada instância de dados de treinamento é inserido no arquivo ARFF pelo script bash por meio de uma lista previamente confeccionada com as classificações. Estes rótulos das instâncias de treinamento podem ser, por exemplo, “Positive”, “Negative” ou “?”. Em que pese, o símbolo “?” é utilizado quando se pretende testar uma proteína não utilizada no treinamento da rede neural; um exemplo de formatação das proteínas da Tabela 1 para o formato ARFF do WEKA está no Anexo II.

Para melhor compreensão do que se trata um esboço visual e simplificado do fluxo de execução da tradução da cadeia de aminoácidos para uma estrutura passível de processamento pelo software WEKA, segue abaixo a Figura 11:

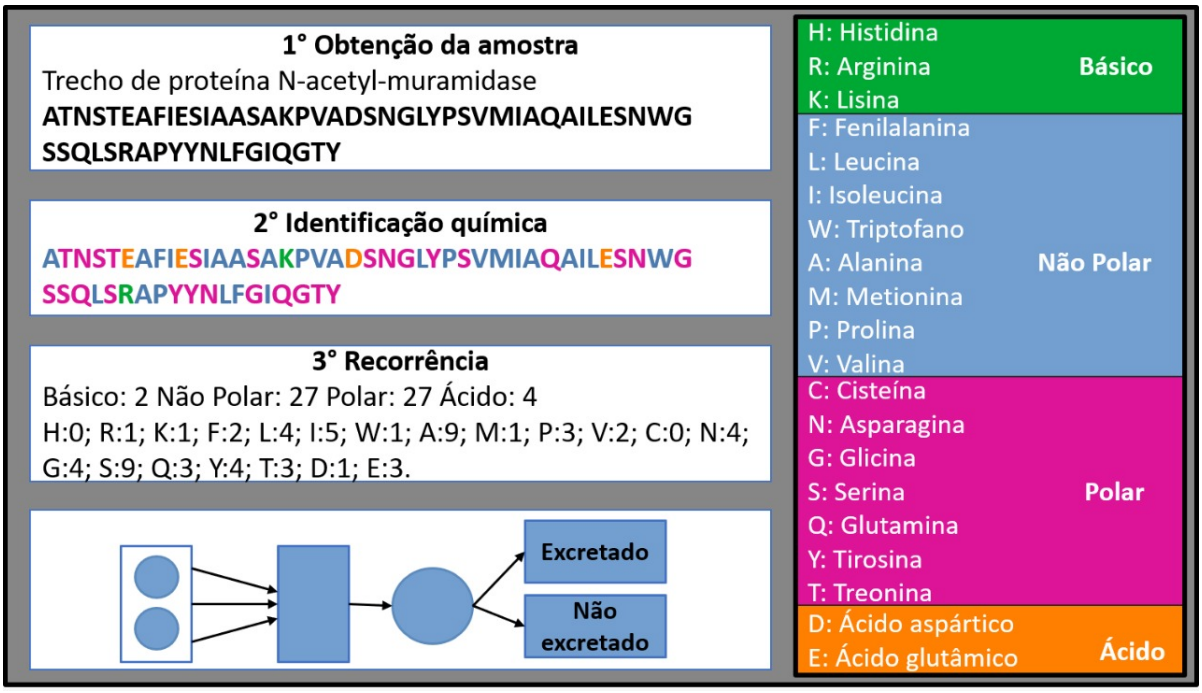

Figura 11 – Esboço da metodologia, onde no primeiro passo obtém as amostras de proteínas, no segundo passo determinas as características e no terceiro converte as características em valores para se treinar a rede neural artificial.

Criada pelo autor

Além das proteínas coletadas na literatura, utilizou-se também o conjunto de proteínas treinadas pelo software PeNGaRoo. Este conjunto de dados foi composto por dois subgrupos: um grupo para treinamento e um para teste. Ambos os grupos contém dois arquivos sendo o conjunto positivo as proteínas secretadas por vias não clássicas e o grupo negativo composto por proteínas secretadas por vias clássicas . Esses arquivos estão em um repositório com o nome de Training Dataset e IndependentDataset no servidor do PeNGaRoo. Também utilizou-se 3150 proteínas com anotações de membrana (integral e parcial) e secretadas por vias clássicas (1050 proteínas por grupamento) para compor o controle negativo. Essas proteínas foram obtidas de uma base com dados classificados dos genomas da Tabela 1. As proteínas classificadas como citoplasmáticas foram excluídas do conjunto de treinamento porque a classificação não considerou a possibilidade de PSVnC's e sabe-se que, atualmente, a maior parte das PSVnC's comprovadas são consideradas citoplasmáticas.

No que tange a rede neural supervisionada, esta foi criada com uma camada interme-

diária com 16 neurônios e 1 neurônio na camada de saída. O número de camadas ocultas foi definida pela fórmula (número de atributos + número de classes) / 2 que é o padrão do *WEKA*. Esta arquitetura foi definida por meio de testes efetuados com outras configurações. Porém, a configuração apresentada foi a que retornou melhores resultados para a rede neural treinada, pois o aumento de camadas e números de neurônios não trouxeram ganhos para a rede.

3.4 Validação e Comparação de Resultados

Essa etapa objetivou validar os passos anteriores dando robustez à pesquisa. Foi efetuada uma comparação com outros preditores de PSVnC's avaliando as mesmas proteínas catalogadas na primeira etapa, via o percentual de assertividade da rede neural. O processo consistiu em tentativa e erro com o intuito de decidir sobre a qualidade de um conjunto de descritores até que, ao menos, mínimos locais fossem alcançados. No que toca a expressão mínimo local, esta atenta para o fato da quantidade de descritores e suas prováveis combinações gerarem uma quantidade proibitiva a ser explorada extensivamente por uma busca sequencial. Por sua vez, a busca por quais descritores produzem o melhor resultado de classificação pode ser o foco de outro projeto utilizando, por exemplo, algoritmos genéticos para explorar uma quantidade de grandeza fatorial de possibilidades de agrupamentos para todos os índices de propensão de aminoácidos presentes no AAindex. Outra forma de avaliação consistiu em estimar a qualidade das classificações do modelo com o treinamento da rede, utilizando de todos os dados coletados na primeira etapa como conjunto de treinamento em que se testou a rede com o conjunto de dados do software PeNGaRoo e ainda, para validação da rede utilizou-se as proteínas obtidas no trabalho *Common Non-classically Secreted Bacterial Proteins with Experimental Evidence* (WANG et al., 2016). A utilização de dados que não participaram do treinamento do modelo ajudou a evitar o super ajuste do modelo aos dados, que consiste em quando a rede já conhece o valor de saída da amostra submetida e já é capaz de predizer resultados. Isso ocorre quando a mesma amostra é utilizada tanto no treinamento quanto na validação.

Capítulo 4

Experimentos e Análise dos Resultados

No decorrer deste capítulo serão apresentados os experimentos realizados utilizando as técnicas descritas na metodologia. Nele encontra-se cada etapa dos testes efetuados nas características propostas para o treinamento da rede neural e por último realiza-se uma comparação de desempenho com outros preditores e uma análise dos resultados obtidos.

4.1 Conjunto de dados, treinamento e validação

Uma etapa fundamental desta pesquisa consistiu em obter conjuntos de amostras significativas de proteínas para diferentes tipos de secreção proteica. Após obtido, esses conjuntos foram divididos em dois grupos, quais sejam: grupo positivo e negativo, a fim de colaborar para um eficiente treinamento e validação da rede treinada. Entretanto, cabe ressaltar que as proteínas que compuseram o grupo positivo (proteínas de secreção por vias não clássicas) são escassas na literatura atual e por tanto, mostrou-se necessário o detalhamento do conjunto de proteínas que compuseram o conjunto de dados de treinamento e validação.

4.1.1 Conjunto de dados de treinamento

O conjunto real de dados utilizados para o treinamento foi obtido do artigo (ZHANG et al., 2020) que descreve o preditor *"PeNGaRoo"* que, disponibiliza bases de proteínas positivas e negativas para treinamento e validação. A base de treinamento denominada de *Training Dataset* é composta por 587 proteínas sendo 446 proteínas negativas (não secretadas por vias não clássicas) e 141 proteínas positivas (proteínas secretadas por vias não clássicas). Também foram utilizadas proteínas de anotações SEC obtidas do repositório UNIPROT para compor o conjunto de proteínas negativas para o treinamento.

Para avaliar o desempenho dos índices de propensão no treinamento da rede foi utilizada a base de proteínas a qual não foi utilizada pra treinamento, denominada *Independent Dataset* que é composta por 34 proteínas positivas e 34 proteínas negativas.

4.1.2 Conjunto de dados de validação

No que diz respeito ao conjunto de dados de validação, uma base foi montada utilizando o artigo (WANG et al., 2016) de onde foi possível obter 43 proteínas secretadas por vias não clássicas, dentre este grupo foi criado um subgrupo de 13 proteínas da bactéria *Bacillus subtilis (b.Subtilis)*. As estruturas das proteínas foram retiradas do site [1] e serviram de suporte para a montagem de um arquivo de validação positiva, enquanto para a validação negativa foi utilizado um conjunto de proteínas de anotações de membrana (integral e parcial) e um conjunto de anotação SEC secretadas por vias clássicas.

4.2 Análise Experimental

Todos os experimentos descritos a seguir foram executados seguindo o método proposto no capítulo 3 deste trabalho. Inicialmente, todos os índices de propensão encontrados durante a pesquisa foram submetidos para teste e validação e, para o caso de um resultado não satisfatório, repetia-se todo o processo com o intuito de melhorar os resultados. Dessa forma, foram obtidos três conjuntos de índice de propensão que foram organizados de acordo com a leitura de trabalhos relacionados e coleta de dados, sendo o grupo 3 montado com a junção do grupo 1 e 2, permitindo um grupo com uma quantidade maior de características para melhor treinar a rede neural artificial. Deste modo, os grupos são descritos a seguir:

- **Grupo 1**: BASIC, ACID, POLAR, NONPOLAR, MAS, CHAM830103, CHAM830104, CHAM830105, FAUJ880111, MONM990201, MONM990101.
- **Grupo 2**: BASIC, ACID, POLAR, NONPOLAR, BEGF750101, BROC820102, FAUJ880112, GEIM800103, GEIM800105, LEWP710101, NAKH900102, NAKH900108, OOBM850104, PALJ810115, PONP800106, QIAN880116, RICJ880107, ROBB760111, ROSM880103, VENT840101, AURR980101, AURR980105, AURR980118, ZHOH040103.
- **Grupo 3**: BASIC, ACID, POLAR, NONPOLAR, CHAM830103, CHAM830104, CHAM830105, FAUJ880111, MONM990201, MONM990101, BEGF750101, BROC820102, FAUJ880112, GEIM800103, GEIM800105, LEWP710101, NAKH900102, NAKH900108, OOBM850104, PALJ810115, PONP800106, QIAN880116, RICJ880107, ROBB760111, ROSM880103, VENT840101, AURR980101, AURR980105, AURR980118, ZHOH040103.

Todos os três conjuntos têm em comum a frequência de repetição de cada aminoácido contido na proteína. Características estas calculadas implicitamente dentro do algoritmo *features*, conforme descrito na seção 3.3.

[1] Disponivel em: <https://www.uniprot.org/> Acessado em 20 dezembro 2020

Neste sentido, a Figura 12 esboça o fluxo de como foi efetuado a escolha do melhor grupo de características, em que foi treinado três redes, cada uma com os respectivos grupos 1, 2 e 3, mas com o mesmo conjunto de proteína. Quanto a validação do treinamento, este se deu com a submissão de proteínas que não participaram do treinamento, validando assim, por meio de porcentagem de assertividade possibilitando determinar o melhor grupo de características.

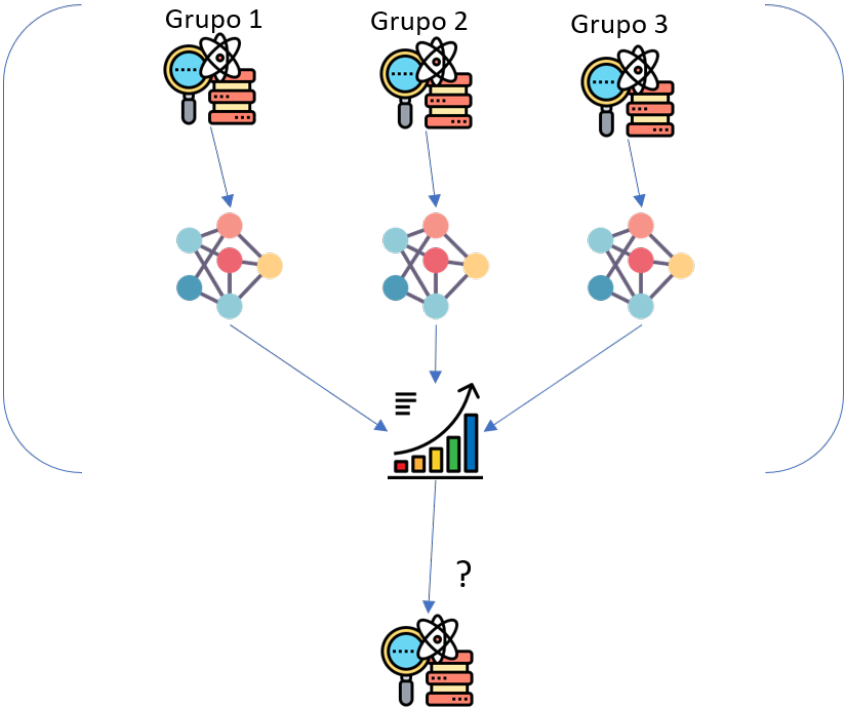

Figura 12 – Representação de como foi determinado o melhor grupo de características para treinamento da rede neural artificial.

Uma vez definido o grupo de índice de propensão a ser testado, o mesmo, juntamente com o conjunto de proteínas de treinamento foi submetido para o *valifasta*, que retornou o arquivo formatado para ser submetido como um arquivo de entrada no Software Weka. Representando a matriz de confusão (valores reais preditos pelo preditor) para os grupos 1, 2 e 3, segue a Tabela 2:

Tabela 2 – Representação da matriz de confusão do teste dos três grupos de índice de propensão testados.

Matriz de Confusão

VN 181	FN 45	
FP 35	VP 106	
Grupo 1		

VN 229	FN 16	
FP 22	VP 119	
Grupo 2		

VN 414	FN 30	
FP 39	VP 102	
Grupo 3		

Quanto aos resultados das medições de desempenho do treinamento da rede neural para cada um dos três grupos de índice de propensão, segue a Tabela 3:

Tabela 3 – Resultados da sensibilidade, Especificidade, Acurácia, Valor preditivo positivo (VPP), Valor preditivo negativo (VPN) de cada grupo testado.

Teste	**Sensibilidade**	**Especificidade**	**Acurácia**	**(VPP)**	**(VPN)**
Grupo 1	0,70	0,83	0,78	0,75	0,80
Grupo 2	0,88	0,91	0,97	0,84	0,93
Grupo 3	0,77	0,91	0,88	0,72	0,93

Após o treinamento e obtenção dos dados de validação, as redes treinadas foram submetidas para uma validação utilizando-se o conjunto de dados *Independent Dataset*. Cabe salientar que, o objetivo aqui é verificar se as redes de fato apresentaram resultados verídicos e não um possível sobreajuste. Assim, o gráfico contido na Figura 13 demonstra os resultados obtidos, pois trata-se de gráfico de barras da imagem que demonstra a porcentagem de acerto nos conjuntos positivos e negativos testados.

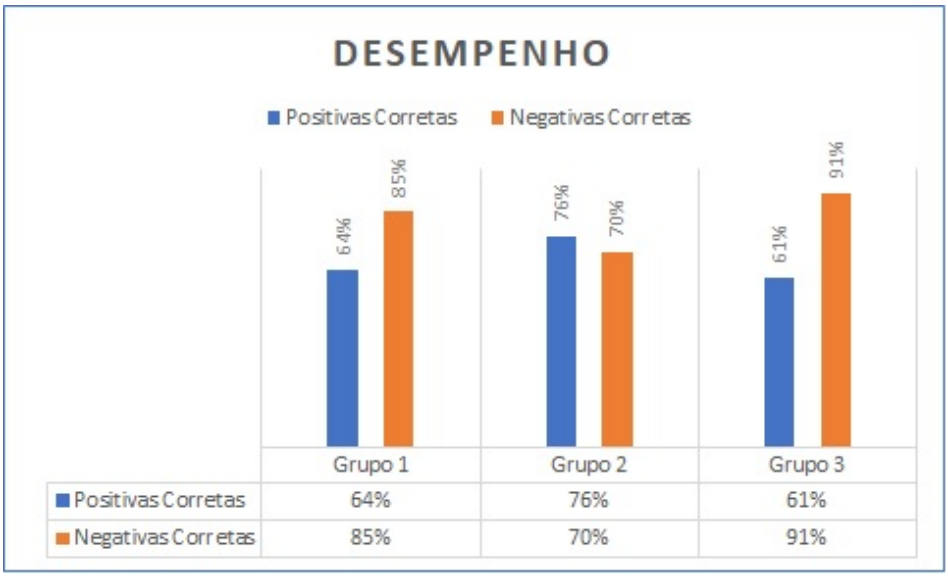

Figura 13 – Representação dos resultados da submissão do dataset Independent Dataset.

Deste modo, abragendo a análise a considerar também a Tabela 3, os resultados indicam o grupo 2 como o possível melhor conjunto de índice de propensão a apresentar os melhores valores de Sensibilidade, Especificidade, Acurácia, VPP e VPN. Por outro lado, se analisarmos os resultados do gráfico contido na Figura 13 isoladamente, estes valores não se monstram reais, no sentido de que o grupo 2 obteve um resultado discrepante do seu treinamento, consultando a Tabela 3, em que pese, a indicação de melhor precisão em proteínas consideradas negativas com uma porcentagem de 91% de especificidade. No

entanto, o gráfico da Figura 13 indica 70% de predição correta para proteínas que não são secretadas e 76% para as secretadas; resultado diferente do esperado da Tabela 3. Desta forma o grupo 3 foi escolhido para a sequência de testes pelo fato de apresentar um valor real condizente com os valores apresentados na Tabela 3 e valores de desempenho Figura 13, indicando uma porcentagem de especificidade igual a Tabela 3. Outro aspecto importante se deu pelo fato de o grupo 3 ter apresentado maior valor de predição correto para proteínas negativas, que não são de secreção por vias não clássicas. Salienta-se que, esta ocorrência é importante, valendo-se da importância de não se criar um preditor tendencioso a falso positivo.

Para melhorar o desempenho de predição positiva foi efetuado um novo treinamento da rede com o grupo 3 de *features*, grupo melhor analisado na etapa anterior porém com o diferencial da modificação da quantidade de proteínas no treinamento.

Deste modo, criou-se uma nova representação na Figura 14, para demonstrar a etapa de determinação da melhor configuração da quantidade de proteínas positivas e negativas com o objetivo de alcançar melhor assertividade.

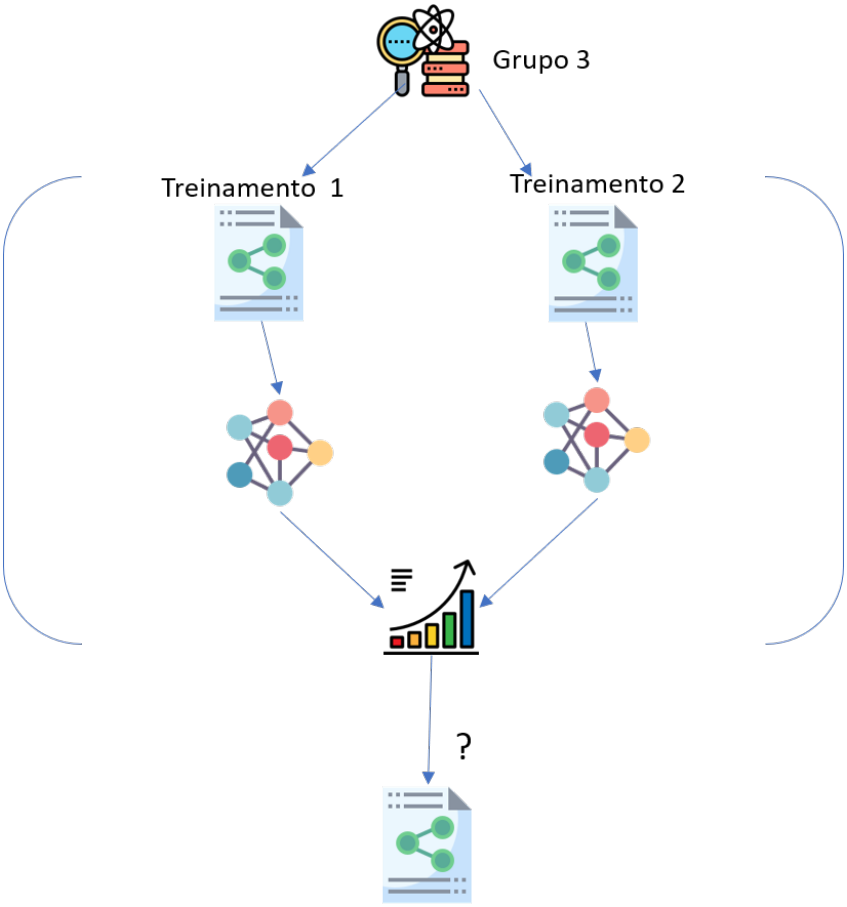

Figura 14 – Representação dos passos para se determinar a melhor quantidade de proteínas no grupo positivo e negativo.

Assim sendo, a rede foi treinada por duas vezes e, em ambos os testes o conjunto de

entrada positiva se manteve igual apenas variando o conjunto de proteínas negativas, em que os resultados foram avaliados com métricas de porcentagem apresentados em gráficos. Segue a divisão dos testes:

- ❏ **Treinamento 1**: 141 proteínas positivas e 141 proteínas negativas.
- ❏ **Treinamento 2**: 141 proteínas positivas e 211 proteínas negativas.

Dito isto, a Tabela 4 dispõe os resultados após o retreinamento da rede com o grupo 3 de índice de propensão.

Tabela 4 – Representação da matriz de confusão do retreinamento do grupo 3 de índice de propensão.

Matriz de Confusão

VN 112	FN 29
FP 29	VP 112

Treinamento 1

VN 174	FN 37
FP 26	VP 115

Treinamento 2

Também foi criado a Tabela 5 para apresenta os valores de sensibilidade, especificidade e acurácia.

Tabela 5 – Resultados da sensibilidade, Especificidade, Acurácia, Valor preditivo positivo (VPP), Valor preditivo negativo (VPN) do retreinamento do grupo 3 de índice de propensão.

Teste	**Sensibilidade**	**Especificidade**	**Acurácia**	**(VPP)**	**(VPN)**
Treinamento 1	0,79	0,79	0,79	0,79	0.79
Treinamento 2	0,75	0,87	0,82	0,81	0.82

Em seguida, as redes foram treinadas e submetidas a um teste de validação com proteínas diferentes daquelas utilizadas no treinamento. Assim sendo, segue o resultado do teste integrado na representação gráfica da Figura 15.

Ao se analisar os resultados do treinamento seguindo a metodologia proposta e, utilizando-se o grupo 3, verifica-se que o treinamento 2 obteve um melhor desempenho se comparado ao 1. Deste modo, a considerar a Tabela 5 é possível notar que 1 obteve uma sensibilidade maior que 2. No entanto, ao analisar os demais índices de especificidade, acurácia, VPP e VPN foi o treinamento 2 que se sobressaiu sobre o 1. Neste sentido, temos que estes índices são comprovados no teste integrado demonstrado no resultado da Figura 15 em que é possível notar que 2 tem um melhor balanceamento de assertividade que o 1, salientando a importância deste tipo de análise com vistas a garantir o treinamento de uma rede que não seja tendenciosa, mas idealmente neutra.

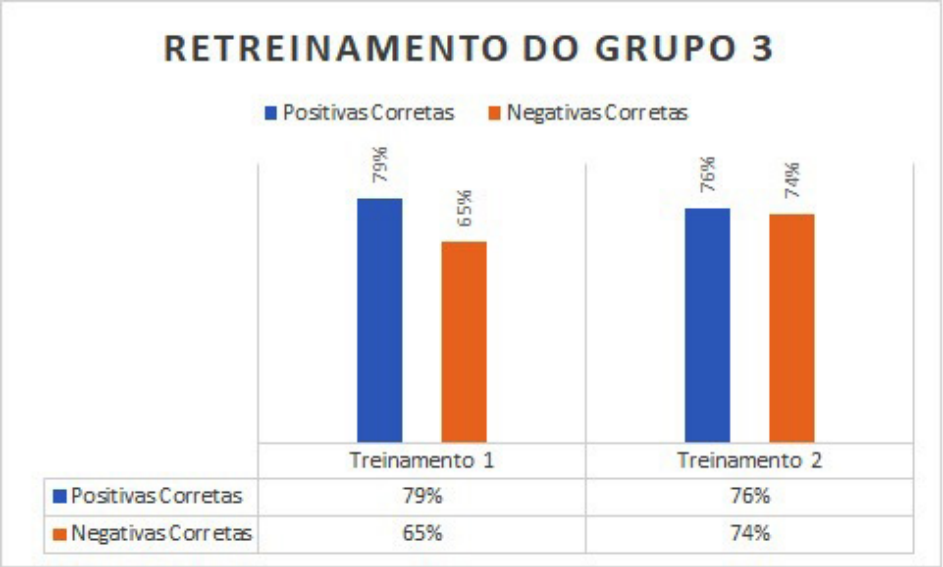

Figura 15 – Representação dos resultados da submissão do dataset Independent Dataset no retreinamento do grupo 3, este gráfico demonstra a porcentagem de acertos nos conjuntos positivos e negativos testados.

Após a obtenção das características ideais para treinamento e a quantidade de proteínas de entrada, a rede (treinamento 2) foi submetida a um teste para validar sua real eficiência, objetivando utilizar esta configuração para a próxima etapa de testes. Posto isto, Figura 16 propõe o desenho de cada um dos passos tomados nesta etapa.

Com isso foi possível avaliar a eficiência das redes treinadas com a utilização de métricas de porcentagem de assertividade positivas e negativas. Segue a divisão das proteínas em grupos que compuseram esta etapa:

Positivo : Proteínas de anotação de secreção não clássica.
Negativas: Proteínas de anotação SEC e membrana.

Avaliando os resultados da rede após o teste, foi identificada que a rede treinada, utilizando apenas o conjunto "Training Dataset", não apresentou um bom desempenho para proteínas de anotação SEC. Por isso, foi necessário refazer o treinamento utilizando outro conjunto de proteínas para compor o conjunto negativo, obtendo assim duas redes treinadas, denominadas de Preditor 1 e Preditor 2 , assim como demonstra a Figura 16, pelo detalhamento dos arquivos que participaram do treinamento:

1. Preditor 1.

   ❏ Entrada positiva: conjunto de proteínas positivas contidas no arquivo "Training

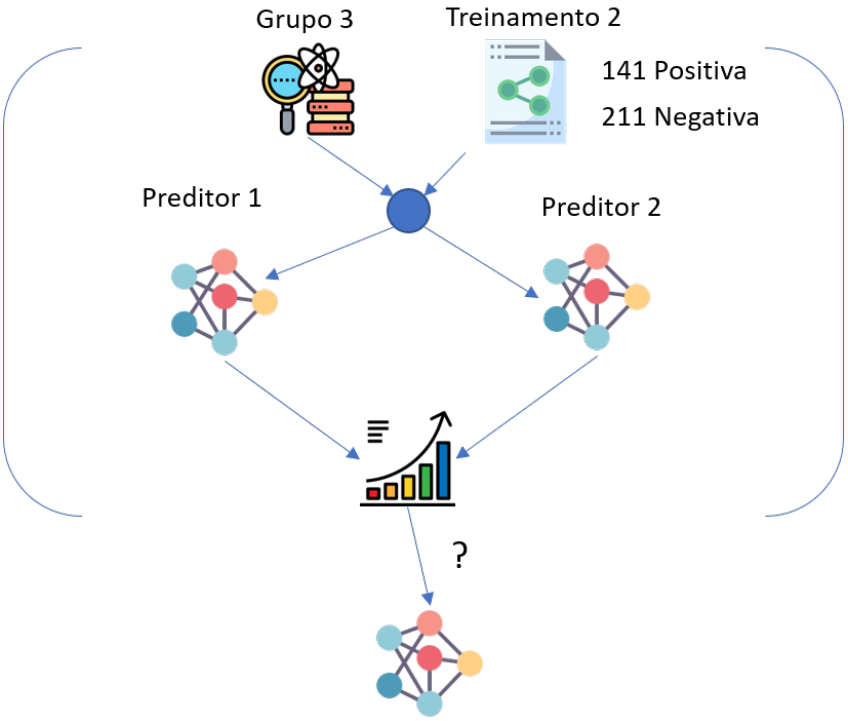

Figura 16 – Representação da etapa que determina o grupo de proteína que apresentou melhores resultados no treinamento.

Dataset“.

- ❏ Entrada negativa - conjunto de proteínas negativas contidas no arquivo “Training Dataset “.

2. Preditor 2.

- ❏ Entrada positiva: conjunto de proteínas positivas contidas no arquivo “Training Dataset“.
- ❏ Entrada negativa: conjunto de proteínas com anotação de secreção SEC.

Salienta-se que, ambas seguiram a mesma proporcionalidade de elemento para compor o treinamento de 141 proteínas positivas e 211 proteínas negativas. Desta maneira, segue o resultado do treinamento efetuado com a ferramenta *Weka*:

Preditor 1 : Correctly Classified Instances 82.1023%.
Preditor 2 : Correctly Classified Instances 92.3295%.

O resultado do treinamento do preditor 1 e 2 são evidenciados na Tabela 6.

Tabela 6 – Resultado do Treinamento entre o Preditor 1 e Preditor 2.

Teste	**Sensibilidade**	**Especificidade**	**Acurácia**	**(VPP)**	**(VPN)**
Preditor 1	0.75	0.87	0,82	0.81	0.82
Preditor 2	0.90	0.93	0.92	0.90	0.93

Quanto a comparação de desempenho entre os preditores, estes serão descritos no próximo capítulo com a submissão de um teste de classificação.

4.3 Avaliação dos Resultados

A fim de avaliar o método proposto, foram efetuados dois estudos de comparação. O primeiro, utilizando os Preditores 1 e 2, treinados com o grupo 3 de índice de propensão descrito na unidade 4.2, em que ambos os preditores 1 e 2 foram submetidos a arquivos de teste a fim de se obter o melhor preditor. E o segundo estudo de comparação foi realizado utilizando dois outros preditores de secreção por vias não clássicas, *SecretomeP* e *PeNGaRoo*, e ainda, o melhor preditor obtido do primeiro resultado de comparação; resultados apresentados em 4.3.1. Posto isto, todos os três preditores foram submetidos a uma nova predição utilizando o conjunto de validação descrito na unidade 4.1.2 de título Conjunto de dados de validação.

4.3.1 Comparação do Preditor 1 com o Preditor 2 com submissão de conjunto de dados

Nesta validação ambos os preditores treinados e descritos anteriormente receberam o mesmo conjunto de entrada de teste. Foi utilizado como conjunto positivo as proteínas positivas contidas no arquivo "Independent Dataset" e como proteínas negativas foram utilizadas proteínas de anotações de secreção de membrana e SEC. Deste modo, os resultados são evidenciados na Tabela 7, mostrando a quantidade de acertos e erros entre os preditores.

Tabela 7 – Resultado do Treinamento entre o Preditor 1 e Preditor 2.

Anotação Proteína	**Porcentagem Acertos Preditor 1**	**Porcentagem Acertos Preditor 2**
Conjunto positivo via não clássica	76,47%	73,52%
SEC	46,66%	82,22%
Membrana	100%	92,85%

Na Tabela 7 é possível notar que o Preditor 1 tem um desempenho ligeiramente superior ao Preditor 2 nas proteínas positivas (via não clássica) e membranas, porém, de-

monstra um pior resultado para predição de proteínas de anotações SEC. Com relação a Tabela 6, nota-se que Preditor 2 possui um valor de acurácia superior ao Preditor 1, bem como, é possível observar que ele possui desempenho mais equilibrado entre todos os tipos de proteínas submetidas ao teste. Isto posto, com base nessas análises justifica-se a escolha do Preditor 2 para participar do próximo estudo de comparação.

4.3.2 Comparação de preditores com submissão conjunto de dados de validação

O segundo teste foi realizado para avaliar a eficiência dos índices de propensão e do conjunto de proteínas de treinamento obtidos pela metodologia proposta neste trabalho.

Para tanto, submeteu-se um total de 13 proteínas secretadas por vias não clássicas da bactéria (*b.Subtilis*), 13 de anotação de membrana e 13 proteínas SEC no preditor treinado, proteínas já detalhadas na unidade 4.1.2 denominada Conjunto de dados de validação. Nesta etapa o resultado da predição do Preditor 2 foi comparado com os resultados de outros dois preditores SecretomeP e PeNGaRoo. Neste sentido, vale salientar que, considera-se o sucesso da rede no seu treinamento quando os valores preditos são próximos ou superiores à predição dos demais preditores utilizados para comparação. Sendo assim, para a representação do fluxo desta etapa, segue Figura 17:

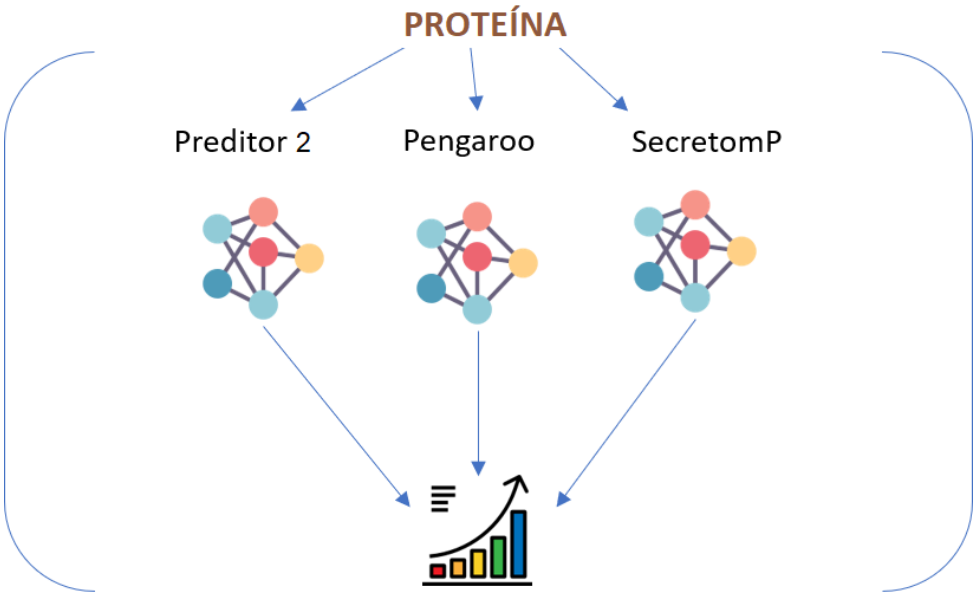

Figura 17 – Representação da etapa onde se efetua o teste de comparação do proditor 2 com outros preditores.

Para melhor esplanação, as Tabelas 8, 9 e 10 expõem os resultados da predição da

rede treinada em que se utilizou o método proposto, comparando-os com dois outros preditores. Observa-se em ambas as tabelas o número total de amostras positivas e negativas submetidas e a quantidade de acertos entre os preditores.

Dada a Tabela 8, observa-se o primeiro resultado, apresentando a eficiência dos preditores para proteínas consideradas secretadas por via não clássicas .

Tabela 8 – Informações obtidas através da submissão de proteínas de bactéria *b.Subtilis* denominadas cientificamente como secreção por vias não clássicas para os três preditores descritos.

Não Clássicas b.subtilis	**Total de amostra**	**Predição Correta**	**Predição Incorreta**
Preditor 2	13	13	0
Pengaroo	13	13	0
Secretomep 2.0	13	2	11

O segundo resultado trata da eficiência dos preditores para proteínas consideradas de membranas que são consideradas como não secretadas representados na Tabela 9.

Tabela 9 – Informações obtidas através de submissão de proteínas de anotação de membranas, proteínas que não são exportadas para o meio extracelular para os três preditores descritos.

Membranas	**Total de amostra**	**Predição Correta**	**Predição Incorreta**
Preditor 2	13	12	1
Pengaroo	13	10	3
Secretomep 2.0	13	0	13

O terceiro e último resultado, expõe na Tabela 10, a eficiência dos preditores para proteínas de anotação SEC que são proteínas consideradas não secretadas por vias não clássicas.

Tabela 10 – Informações obtidas através de submissão de proteínas de anotação de secreção por vias clássicas, proteínas que são exportadas para o meio extracelular por vias conhecidas.

Sec	**Total de amostra**	**Predição Correta**	**Predição Incorreta**
Preditor 2	13	9	4
Pengaroo	13	3	10
Secretomep 2.0	13	9	4

Com o objetivo de melhor expor os resultados, a Figura 18 esboça um gráfico contendo a porcentagem de acertos positivos entre os três preditores em que, é possível visualizar a performance de assertividade para cada um dos preditores em diferentes tipos de proteínas submetidas.

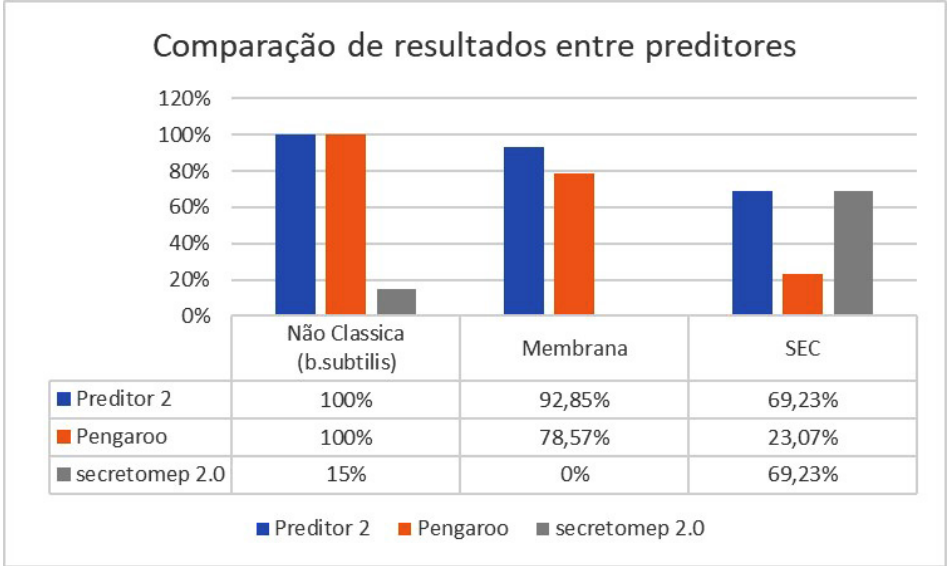

Figura 18 – Comparação de acerto entre o preditor treinado (Preditor 2) e os preditores Pengarro e Secretomep 2.0 utilizando proteínas de tipo Não Clássica, proteínas de Membranas e Sec.

A considerar estes resultados, é possível notar que o conjunto de índice de propensão e o conjunto de proteínas utilizadas para treinar a rede contêm uma boa correlação entre os dados, principalmente quando se compara os resultados do preditor *PeNGaRoo*. Ambos têm a mesma porcentagem de acerto para as proteínas da bactéria *B.subtilis*, uma vez que foi comprovado em laboratório que estas proteínas são de fato secretadas por vias não convencionais. Cabe ressaltar que o preditor *PeNGaRoo*, uma vez que a proteína já é conhecida no meio científico como secretada por vias não clássica, não ocorre a execução do algoritmo que prediz a amostra enviada, apenas a classifica como secretada por vias não classicas. Este fato ocorre com estas amostras da bactéria *b.subtilis*, mas com as demais amostras enviadas de Membrana e Sec ele efetua a predição. Este fato, justifica o porquê de uma eficiência inferior na predição dessas amostras citadas acima se comparadas às amostras de secreção por vias não clássica. No entanto, a rede treinada pelo método proposto efetua uma predição em todos os tipos de proteínas, independentemente se ela já é conhecida ou não.

Assim, o Preditor 2 apresenta um bom desempenho em relação às proteínas de secreção por vias não clássicas, Membranas e Sec. Apesar de ter predito todas as proteínas de secreção por vias não clássica como positivo, isto não significa que ele possua sensibilidade

de 1, pois na realidade sua sensibilidade é 0,90 e especificidade de 0,93 (Tabela 6).

O teste foi efetuado com 13 proteínas que, apesar de serem comprovadamente secretadas por vias não clássicas, são apenas uma pequena amostra de uma vasta quantidade de proteínas existentes com essas características ainda desconhecidas do meio científico, este fato dificulta a medição real da eficiência do preditor. Desta forma, é necessário cautela ao afirmar que o preditor treinado é totalmente assertivo para as proteínas positivas, pois ele apresenta 92,33% de eficiência no conjunto total e tem um desempenho superior ou equivalente aos outros preditores para proteínas de secreção não clássica, Sec e proteínas de membrana.

Afim de esboçar melhor os resultados apresentados pelo preditor treinado, foi efetuada uma nova submissão de amostras de proteínas apenas para o Preditor 2, dividindo-as em três grupos, quais sejam: secretadas por vias não clássica, SEC e proteínas de validação utilizou-se conjuntos com 43 proteínas para cada grupo, cujos dados foram detalhados na seção 4.1.2 (Conjunto de dados de validação). Segue a amostragem na Tabela 11.

Tabela 11 – Informações obtidas através de submissão de proteínas de anotação de secreção por vias não clássicas, Sec e membrana para validar a eficiência do Preditor 2.

Preditor 2	**Total de amostra**	**Predição Correta**	**%**
P. Via não Classica	43	41	95,34%
SEC	43	37	86,04%
Membrana	43	39	91%

Assim, com base na Tabela 11, é possível notar que o Preditor 2 de fato apresenta resultados coerentes com os testes efetuados entre os demais preditores, pois obteve bons resultados para os diferentes tipos de proteínas e uma melhor precisão nas proteínas secretadas por vias não clássicas, verificando a eficácia e a relevância do Preditor 2, treinado pelo método proposto e corroborando o objetivo deste trabalho.

4.3.3 Conclusões

Os experimentos aqui descritos possibilitaram avaliar as etapas previstas pela metodologia, sendo elas: avaliação dos conjuntos de características de treinamento; avaliação dos conjuntos de amostras das proteínas de treinamento e validação; consideração da natureza de secreção; quantificação ótima das proteínas para compor os conjuntos de treinamento. Os resultados apresentados evidenciaram que a utilização de 141 proteínas no conjunto de amostras positivas e 211 nas amostras negativas no grupo 3 de características proporcionou resultados superiores na predição de proteínas secretadas por vias não clássicas quando comparados a preditores descritos na literatura correlata.

Capítulo 5

Discussões Finais

O objetivo deste trabalho foi encontrar características físicas, químicas e biológicas capazes de classificar se uma proteína bacteriana é secretada por vias não clássicas utilizando redes neurais. Para alcançar este objetivo foi necessário efetuar pesquisas em bibliografias relacionadas, buscando essas características, bem como, conjuntos de proteínas sabidamente positivas e negativas para treinar e validar a rede neural.

Uma vez obtido o melhor conjunto de índice de propensão a partir da análise dos resultados preditos pelas redes treinadas, tal conjunto foi usado para validar a metodologia aplicada. Os experimentos também avaliaram a influência do tipo de proteína (SEC, membrana e secretadas por vias não clássicas) e da quantidade de elementos que compunham os arquivos de treinamento para se treinar uma rede e, dessa forma, obter a eficiência da rede em predizer resultados coerentes. Os resultados obtidos evidenciaram que o grupo de propensão denominado grupo 3 se sobressaiu comparado aos demais índices utilizados com resultados satisfatórios e que contribuíram para treinamento da rede neural. Ficou evidente também que, a quantidade de proteínas utilizadas nos conjuntos de treinamento pode influenciar no resultado final da rede. Quanto aos testes efetuados variando a quantidade de elementos que compunham o arquivo de treinamento negativo, concluiu-se que, um conjunto dataset (Positivo, Negativo) com proporções idênticas não foi eficiente para proporcionar bons resultados para a rede. Já os testes que utilizaram uma proporção maior de elementos no conjunto de treinamento negativo proporcionou resultados melhores para o treinamento da rede. Assim sendo, deve-se destacar também, que a utilização do dataset Training Dataset (*Pengaroo*) somente com o conjunto positivo e negativo obtiveram resultados satisfatórios, porém, inferiores para proteínas de anotação SEC. Dado este fato, deu-se novo treinamento utilizando somente proteínas positivas do arquivo Training Dataset (*Pengaroo*) com um conjunto negativo montado, composto por proteínas de anotação SEC e, neste novo treinamento resultados se mostraram melhores. Vale ressaltar que proteínas de anotação de secreção de membranas não foram utilizadas em nenhum dos grupos de treinamento, para que a rede conseguisse identificar um padrão utilizando-se apenas de proteínas contidas no grupo positivo e negativo.

Para validar este experimento foram utilizados conjuntos de proteínas cuja natureza de secreção já era conhecida. Assim, a análise da precisão de classificação da rede neural treinada neste estudo (Preditor 2) foi feita comparando-se o conjunto de saída da mesma com dois outros preditores, também já conhecidos no meio científico. Os resultados evidenciaram que o Preditor 2 teve resultados superiores ou idênticos na predição correta de proteínas secretadas por vias não clássicas.

O conjunto total dos experimentos obtiveram um resultado satisfatório, uma vez que foi possível obter uma rede equilibrada e um bom desempenho em classificar corretamente as proteínas de secreção por vias não clássicas como secretadas e proteínas de membrana e Sec como não secretadas.

Um grande desafio encontrado durante a realização deste trabalho foi reunir proteínas já conhecidas como secretadas por vias não clássicas, assim como características físicas químicas e biológicas dos aminoácidos. Esta dificuldade se deve ao fato de não haver um consenso sobre o fenômeno de secreção por vias não clássicas e de não ser claro quais características de uma proteína evidenciam que ela será secretada por essa via. Os resultados satisfatórios da metodologia utilizada neste trabalho são evidenciados nos testes cujo conjunto de índice de propensão utilizado obteve bons índices de sensibilidade e especificidade, bem como, na comparação destes resultados com outros preditores.

5.1 Conclusão

A partir dos resultados apresentados neste trabalho é possível afirmar que o método proposto atingiu seu objetivo, demonstrando combinação de determinados índices de propensão e um bom balanceamento de amostras no conjunto de treinamento, possibilitando obter um classificador utilizando redes neurais, método relativamente eficaz na classificação de proteínas secretadas por vias não clássicas.

5.2 Principais Contribuições

Dado aos testes e resultados apresentados neste trabalho, vê-se a possibilidade da contribuição acadêmica, no que se refere a pesquisadores que buscam estudar quanto a hipótese da secreta de proteínas por vias não clássicas para desenvolver suas pesquisas. Além disso, este trabalho possibilitará a viabilização de pesquisas que aperfeiçoem e criem novas técnicas de predição de secreção proteica até mesmo em outras áreas relacionadas.

5.3 Trabalhos Futuros

Com o avanço da tecnologia e dos novos resultados em pesquisas biológicas é sugerido compor um conjunto relativamente grande de proteínas de secreção por vias não clássicas

para um novo treinamento da rede utilizando os mesmos índices de propensão. Outra sugestão possível trata-se da utilização de algoritmos genéticos para testar todos os índices de propensão contidos no repositório AAindex em conjunto com uma rede neural ou outro tipo de classificador.

Referências

AAINDEX. **AAindex**. [S.l.], 2020. Disponível em: <http://www.genome.ad.jp/aaindex>. Acesso em: 22 dezembro. 2020.

BAILEY, T.; HARRIS, I. Microhardness studies of a nd-fe-b permanent magnet alloy. **Journal of materials science letters**, Springer, v. 4, n. 2, p. 151–153, 1985.

BARRETO, G. d. A. **Redes neurais não-supervisionadas para processamento de sequências temporais**. 152 p. Tese (Doutorado) — Universidade de São Paulo, 1998.

BATISTA, G. E. d. A. P. et al. **Pré-processamento de dados em aprendizado de máquina supervisionado**. 232 p. Tese (Doutorado) — Universidade de São Paulo, 2003.

BENAVIDES, A. R. d. V. Curvas roc (receiver-operating-characteristic) y sus aplicaciones. 2017. Trabalho de Conclusão de Curso - Universidad de Sevilla.

BENDTSEN, J. D. et al. Non-classical protein secretion in bacteria. **BMC microbiology**, BioMed Central, v. 5, n. 1, p. 1–13, 2005.

BORATYN, G. M. et al. Blast: a more efficient report with usability improvements. **Nucleic acids research**, Oxford University Press, v. 41, n. W1, p. W29–W33, 2013.

BRAGA, A. de P.; FERREIRA, A. C. P. de L.; LUDERMIR, T. B. **Redes neurais artificiais: teoria e aplicações**. [S.l.]: LTC Editora Rio de Janeiro, Brazil:, 2007. 900 p.

CARMINATI, R. et al. Determinação da sensibilidade e da especificidade de um teste de elisa indireto para o diagnóstico de linfadenite caseosa em caprinos. **Revista de Ciências Médicas e Biológicas**, v. 2, n. 1, p. 88–93, 2003.

CHEN, J. et al. A novel strategy for protein production using non-classical secretion pathway in bacillus subtilis. **Microbial cell factories**, Springer, v. 15, n. 1, p. 69, 2016.

COELHO, T. A. **Classificação de Proteínas Com Redes Neurais Artificiais**. [S.l.], 2016. Disponível em: <http://repositorio.ufla.br/bitstream/1/5305/2/Artigo_Class_Prot_RNA.pdf>. Acesso em: 22 dezembro. 2020.

COSTA, J. A. F. et al. **Classificação automática e análise de dados por redes neurais auto-organizáveis.** 359 p. Tese (Doutorado) — Unicamp, 1999.

CRISTIANO, M. V. d. M. B. **Sensibilidade e especificidade na curva roc: um caso de estudo**. 126 p. Tese (Doutorado) — Universidade de Medicina, Universidade do Porto, 2017.

ELMEZAYEN, A. D. et al. Drug repurposing for coronavirus (covid-19): in silico screening of known drugs against coronavirus 3cl hydrolase and protease enzymes. **Journal of Biomolecular Structure and Dynamics**, Taylor & Francis, p. 1–13, 2020.

FERNEDA, E. Redes neurais e sua aplicação em sistemas de recuperação de informação. **Ciência da Informação**, SciELO Brasil, v. 35, n. 1, p. 25–30, 2006.

FLECK, L. et al. Redes neurais artificiais: Princípios básicos. **Revista Eletrônica Científica Inovação e Tecnologia**, v. 1, n. 13, p. 47–57, 2016.

FRANK, E. et al. Data mining in bioinformatics using weka. **Bioinformatics**, Oxford University Press, v. 20, n. 15, p. 2479–2481, 2004.

FRONZES, R.; CHRISTIE, P. J.; WAKSMAN, G. The structural biology of type iv secretion systems. **Nature Reviews Microbiology**, Nature Publishing Group, v. 7, n. 10, p. 703–714, 2009.

GREEN, E. R.; MECSAS, J. Bacterial secretion systems: an overview. **Virulence Mechanisms of Bacterial Pathogens**, Wiley Online Library, p. 213–239, 2016.

GUELPELI, M. V.; RIBEIRO, C. H.; OMAR, N. Utilização de aprendizagem por reforço para modelagem autônoma do aprendiz em um tutor inteligente. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2003. v. 1, n. 1, p. 465–474.

GUIMARÃES, A. M. et al. Módulo de validação cruzada para treinamento de redes neurais artificiais com algoritmos backpropagation e resilient propagation. **Publicatio UEPG: Ciências Exatas e da Terra, Agrárias e Engenharias**, v. 14, n. 01, 2008.

HAYKIN, S. **Redes neurais: princípios e prática**. [S.l.]: Bookman Editora, 2007.

HONDA, H.; FACURE, M.; YAOHAO, P. **Os Três Tipos de Aprendizado de Máquina**. Brasília: lamfo, 2017. Disponível em: <https://lamfo-unb.github.io/2017/07/27/tres-tipos-am/>. Acesso em: 22 dezembro. 2020.

HUNG, C.-H. et al. Prediction of non-classical secreted proteins using informative physicochemical properties. **Interdisciplinary Sciences: Computational Life Sciences**, Springer, v. 2, n. 3, p. 263–270, 2010.

JORGE, M. B. Indrodução as redes neurais artificiais. **Florianópolis: Ufsc**, v. 3, 2013.

KANG, Q.; ZHANG, D. Principle and potential applications of the non-classical protein secretory pathway in bacteria. **Applied Microbiology and Biotechnology**, Springer, v. 104, n. 3, p. 953–965, 2020.

KAWASHIMA, S.; KANEHISA, M. Aaindex: amino acid index database. **Nucleic acids research**, Oxford University Press, v. 28, n. 1, p. 374–374, 2000.

KITANI, E. C. **Mapeamento e visualização de dados em alta dimensão com mapas auto-organizados.** 181 p. Tese (Doutorado) — Universidade de São Paulo, SP, Brasil, 2013.

LIGHTGBM. **LightGBM**. [S.l.], 2020. Disponível em: <https://github.com/Microsoft/LightGBM>. Acesso em: 22 dezembro. 2020.

LORENA, A. C.; CARVALHO, A. C. de. Utilizaçao de técnicas inteligentes em bioinformática. **Relatório Técnico**, v. 219, 2003.

MANZAN, J. R. G. et al. **Análise de desempenho de redes neurais artificiais do tipo multilayer perceptron por meio do distanciamento dos pontos do espaço de saída**. 131 p. Tese (Doutorado) — Universidade Federal de Uberlândia, MG, Brasil, 2016.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, Springer, v. 5, n. 4, p. 115–133, 1943.

MERGLER, M. **Welcome To The World Of Peptides**. 2020. Disponível em: <https://www.bachem.com/fileadmin/user_upload/pdf/Catalogs_Brochures/Welcome_to_the_World_of_Peptides.pdf>. Acesso em: 22 dezembro. 2020.

MONARD, M. C.; BARANAUSKAS, J. A. Sistemas inteligentes: fundamentos e aplicações, chapter conceitos sobre aprendizado de máquina. **Rezende (2003)**, v. 15, p. 89–114, 2003.

NETO, A. d. M. et al. **Aprimoramento da anotação N-terminal de proteínas através da predição de peptídeo sinal em proteínas ortólogas e desenvolvimento de uma ferramenta automática para a identificação de grupos ortólogos contendo erros de anotação**. 107 p. Tese (Doutorado) — Fundação Oswaldo Cruz – FIOCRUZ, Belo Horizonte, 2012.

OLIVEIRA, L. S. d. et al. **Perfil do exoproteoma e identificação de proteínas imunogênicas secretadas por Staphylococcus saprophyticus**. Dissertação (Mestrado) — Universidade Federal de Goiás, Goiânia, 2014.

PENGAROO. [S.l.], 2020. Disponível em: <https://pengaroo.erc.monash.edu/>. Acesso em: 22 dezembro. 2020.

PRATI, R.; BATISTA, G.; MONARD, M. Curvas roc para avaliação de classificadores. **Revista IEEE América Latina**, v. 6, n. 2, p. 215–222, 2008.

REZENDE, A. de F. S. et al. In silico identification of corynebacterium pseudotuberculosis antigenic targets and application in immunodiagnosis. **Journal of medical microbiology**, v. 65, p. 521–529, 2016. Disponível em: <https://doi.org/10.1099/jmm.0.000263>. Acesso em: 22 dezembro. 2020.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, American Psychological Association, v. 65, n. 6, p. 386, 1958.

SANTOS, F. R.; ORTEGA, J. M. Bioinformática aplicada à genômica. **Melhoramento Genômico, Minas Gerais: UFV**, p. 93–98, 2003.

SEGATTO, E.; COURY, D. Redes neurais artificiais recorrentes aplicadas na correção de sinais distorcidos pela saturação de transformadores de corrente. **Sba: Controle & Automação Sociedade Brasileira de Automatica**, SciELO Brasil, v. 17, n. 4, p. 424–436, 2006.

SINGH, G.; PANDA, R. K. Daily sediment yield modeling with artificial neural network using 10-fold cross validation method: a small agricultural watershed, kapgari, india. **International Journal of Earth Sciences and Engineering**, v. 4, n. 6, p. 443–450, 2011.

WANG, G. et al. Common non-classically secreted bacterial proteins with experimental evidence. **Current microbiology**, Springer, v. 72, n. 1, p. 102–111, 2016.

ZHANG, Y. et al. Pengaroo, a combined gradient boosting and ensemble learning framework for predicting non-classical secreted proteins. **Bioinformatics**, Oxford University Press, v. 36, n. 3, p. 704–712, 2020.

Glossário

bash é uma das famosas ferramentas de script do Unix e ideal para usuários Linux e administradores de Sistema.

BLAST Basic Local Alignment Search Tool.

clusterização é a tarefa de dividir a população ou os pontos de dados em vários grupos, de modo que os pontos de dados nos mesmos grupos sejam mais semelhantes a outros pontos de dados no mesmo grupo do que os de outros grupos. Em palavras simples, o objetivo é segregar grupos com traços semelhantes e atribuí-los a clusters.

CSV Comma-separated values, são arquivos de texto de formato regulamentado pelo RFC 4180, que faz uma ordenação de bytes ou um formato de terminador de linha, separando valores com vírgulas. Ele comumente é usado em softwares offices, tais como o Microsoft Excel e o LibreOffice Calc.

FASTA é formato padrão utilizado na bioinformática em texto para representar tanto sequencias de nucleotídeos quanto sequencias de peptídeos, no qual os nucleotídeos ou aminoácidos são representados usando códigos de uma única letra. O formato também permite sequências de nomes e comentários precedendo as sequências. A simplicidade do formato FASTA torna mais fácil manipular e analisar sequências usando ferramentas de processamento de texto e linguagens.

grafos conjunto de pontos (vértices) ligados por retas (as arestas). Importante ferramenta da computação na análise, representação e abstração de problemas.

gram-negativas as bactérias Gram-negativas adquirem coloração vermelha quando submetidas a um processo químico denominado coloração de Gram e possuem uma parede celular mais fina que é envolvida por outra membrana externa, portanto os gram-negativos possuem 2 membranas..

gram-positivas as bactérias Gram-positivas adquirem coloração azul quando submetidas a um processo químico denominado coloração de Gram e possuem uma parede celular grossa cobrindo a membrana citoplasmática..

PAAC Pseudo amino acid composition descriptors.

peptídeo sinal sequência sinalizadora constituída de 13 a 36 aminoácidos contida na extremidade amino-terminal de uma proteína. A sequência do peptídeo sinal tem por função marcar as proteínas que serão exportadas para determinados locais, como por exemplo, o ambiente extracelular. Estas proteínas são reconhecidas por meio do peptídeo sinal, o qual, após a exportação, é removido da proteína por meio da ação de proteases.

POSSUM Matriz de Pontuação Específica de Posição (Position-Specific Scoring Matrix).

PSSM Position-Specific Scoring Matrix - Matriz de Pontuação específica de posição, é um tipo de matriz de pontuação usada em pesquisas de proteína BLAST nas quais as pontuações de substituição de aminoácidos são dadas separadamente para cada posição em um alinhamento de seqüência múltipla de proteínas.

Anexos

ANEXO A

Complemento

A.1

Código fonte do programa features.lisp

```
1  ;rlwrap sbcl --dynamic-space-size 1024 --load features.lisp
2  ;(save-lisp-and-die "features" :executable t :save-runtime-
       options t :compression 9 :toplevel 'main)
3  ;(save-lisp-and-die "features" :executable t :save-runtime-
       options t :toplevel 'main)
4  ;The global set of the random state does not work for the
       executable version of the program.
5  ;I will keep it here but changed all random calls to force the
       state update.
6  ;(setq *random-state* (make-random-state t))
7  (defvar *printlist* nil "Vector to store the final result of the
       program features in a format to produce the final output of the
        program")
8  ;Given a text string, a field delimiter and a position this
       function returns the text at the specific location.
9  (defun nth-string (pos tabular delim)
10 (let ( (inicio 0) (fim) (retorno) (max (length tabular)) (cont 0)
       )
11 ;do while not
12 (do ()
13 ((or (< pos 0) (= cont pos) (> inicio max)))
14 (setq fim (position delim tabular :start inicio))
15 (if (not fim) (setq fim max))
16 (setq retorno (subseq tabular inicio fim)
17 inicio (+ fim 1)
18 cont (1+ cont)
```

```
);setq
);do
(if (< cont pos) nil retorno)
);let
);defun
(defun parse-float (s)
(let ((readval (handler-case
(read-from-string s)
(sb-int:simple-reader-error nil)
(end-of-file nil))))
(cond ((realp readval ) (+ readval 0.0))
(t (error (concatenate 'string "not a float: " s))))))
;Returns a list where each component has a pair of text identifier for a fasta sequence
;and correspondent amino acids as a vector.
(defun read-fasta (filename)
(let ((inputfile (open filename :if-does-not-exist nil))
(fastaheader nil)
(filecontent)
(sequence (make-array 10 :fill-pointer 0 :adjustable t))
);let init block
(when inputfile
(loop for line = (read-line inputfile nil) until (not (zerop (length line)))
finally (setq fastaheader line)
);loop to skyp blanklines
(when (char= (elt fastaheader 0) #\>) ;check for fasta format
(loop for line = (read-line inputfile nil)
while line do
(when (> (length line) 0)
(if (char= (elt line 0) #\>)
(progn
(setq filecontent (append filecontent (list (list (subseq fastaheader 1) sequence ))))
(setq fastaheader line sequence (make-array 10 :fill-pointer 0 :adjustable t))
);progn
(loop for aa across line do (vector-push-extend (intern (string-upcase aa)) sequence))
);if found new fasta header
);when there is content
);loop to read aa sequence
```

```
(setq filecontent (append filecontent (list (list (subseq
   fastaheader 1) sequence))))
);when fasta format
(close inputfile)
);when inputfile
filecontent ;return value
);let
);defun
(defun read-propensity (filename)
(let ((inputfile (open filename :if-does-not-exist nil))
(filecontent nil)
(filelabels nil)
);let init block
(when inputfile
(loop for line = (read-line inputfile nil)
while line do
(when (> (length line) 0)
(if (char/= (elt line 0) #\#)
(setf
filelabels (append filelabels
(list (nth-string 1 line #\Space))
);append
filecontent (append filecontent
(list (loop for i from 2 to 21 collect
(parse-float (nth-string i line #\Space))
);loop
);list
);append
);setf
);if
);when there is content
);loop to read number sequence
(close inputfile)
);when inputfile
(values filelabels filecontent) ;return value
);let
);defun
; Collect a list of needle's positions in a haystack
;(all-positions 'G *sequence*)
(defun all-positions (needle haystack)
(loop
for element in haystack
```

```
97  and position from 0
98  when (eql element needle)
99  collect position))
100 ;Determines the size, mean and standard deviation from a
       numerical sample,
101 ;in general, a list of positions of a pattern
102 (defun statistics( data )
103 (if data
104 (let ( (n 0) (mean 0.0) (var 0.0) (diff) )
105 (loop for x across data do
106 (incf n)
107 (setq diff (- x mean)
108 var (+ var (/ (* diff diff (- n 1)) n))
109 mean (+ mean (/ diff n))
110 );setq
111 );loop
112 (if (> n 1)
113 (values n mean (sqrt (/ var (- n 1))) )
114 (values n mean (sqrt (/ var n )) )
115 )
116 );let
117 );if
118 );defun
119 ;Considering that each search pattern is a list of lisp symbols,
       I need to convert the sequence to a string.
120 (defun convert-to-key ( nucseq )
121 (let ( (var "") )
122 (loop for i in nucseq do
123 (setf var (concatenate 'string var (symbol-name i )))
124 )
125 var
126 )
127 )
128 (defun histo-fasta( filename )
129 (let ( (multifastalist (read-fasta filename ))
130 (aatargets (list 'A 'R 'N 'D 'C 'Q 'E 'G 'H 'I 'L 'K 'M 'F 'P 'S
       'T 'W 'Y 'V))
131 (histolist nil)
132 (fastaname)
133 (histogram nil)
134 )
135 (dolist (fastalist multifastalist histogram)
```

```
(setq fastaname (car fastalist)
histolist nil
);setq
(dolist (aa aatargets histolist)
(setq histolist (append histolist (list (count aa (cadr fastalist
   )))))
);inner do
(setq histogram (append histogram (list (list fastaname histolist
   ))))
);outer do
histogram
);let
)
(defun dohistogram( fastalist )
(let ( (aatargets (list 'A 'R 'N 'D 'C 'Q 'E 'G 'H 'I 'L 'K 'M 'F
    'P 'S 'T 'W 'Y 'V))
(histolist nil)
)
(setq histolist nil);setq
(dolist (aa aatargets histolist)
(setq histolist (append histolist (list (count aa fastalist))))
);do
histolist
);let
)
(defun dohistogramini( fastalist )
(let ( (aatargets (list 'A 'R 'N 'D 'C 'Q 'E 'G 'H 'I 'L 'K 'M 'F
    'P 'S 'T 'W 'Y 'V))
(histolist nil)
(size (length fastalist))
(limit 40)
)
(dolist (aa aatargets histolist)
(setq histolist (append histolist (list (count aa fastalist :end
   (if (> size limit) limit size) ))))
);do
histolist
);let
)
(defun dohistogramend( fastalist )
(let ( (aatargets (list 'A 'R 'N 'D 'C 'Q 'E 'G 'H 'I 'L 'K 'M 'F
    'P 'S 'T 'W 'Y 'V))
```

```
172 (histolist nil)
173 (size (- (length fastalist) 60))
174 )
175 (dolist (aa aatargets histolist)
176 (setq histolist (append histolist (list (count aa fastalist :
       start (if (> size 0) size 0) ))))
177 );do
178 histolist
179 );let
180 )
181 (defun dohistogrammid( fastalist )
182 (let ( (aatargets (list 'A 'R 'N 'D 'C 'Q 'E 'G 'H 'I 'L 'K 'M 'F
       'P 'S 'T 'W 'Y 'V))
183 (histolist nil)
184 (size (length fastalist) )
185 (ratio)
186 (start)
187 (end)
188 )
189 (setf ratio (round (/ size 3))
190 start (if (> size ratio) ratio size)
191 end (- size ratio)
192 end (if (> end 0) end size))
193 (dolist (aa aatargets histolist)
194 (setq histolist (append histolist (list (count aa fastalist :
       start start :end end ))))
195 );do
196 histolist
197 );let
198 )
199 ;The main function settles everything for the execution of the
       program.
200 ;This function also is the entry point for the executable created
       from this lisp code.
201 (defun main ()
202 (if (> (length sb-ext:*posix-argv*) 1)
203 (let (
204 ;primeiro par metro    o nome do arquivo
205 (fastafile nil)
206 (propensityfile nil)
207 (searchspace)
208 (milestone)
```

```
(checkmilestone)
(multifasta)
(multifasta-size)
(listof-sequence-names)
(sequence)
(sequence-name)
(sequence-size)
(sequence-number 0)
(sequence-histogram)
(pos)
(histoalphabet '( A L R K N M D F C P Q S E T G W H Y I V ))
(alphabet)
(physicochemicals)
); let pars
(setf fastafile (if (nth 1 sb-ext:*posix-argv*) (nth 1 sb-ext:*posix-argv*) ) );setf
(if (not (probe-file fastafile))
(progn
(format t "~%~a~%" "ERROR: Amino acid fasta file is missing")
(SB-EXT:EXIT)
)
);if
(setf propensityfile (if (nth 2 sb-ext:*posix-argv*)
(nth 2 sb-ext:*posix-argv*)
"propensity.dat")
);setf
(if (not (probe-file propensityfile ))
(progn
(format t "~%~a~%" "ERROR: Propensity data file is missing")
(SB-EXT:EXIT)
)
);if
(multiple-value-bind ( labels props) (read-propensity propensityfile)
(setf
labels (append labels
(loop for place in (list "INI" "END" "MID") append
(loop for label in labels collect (concatenate 'string place label))
)
)
alphabet (append histoalphabet labels)
```

```
248 physicochemicals props
249 multifasta (read-fasta fastafile)
250 multifasta-size (length multifasta)
251 *printlist* (make-array (list (+ multifasta-size 1) (+ (length
       alphabet) 1) ) :initial-element nil)
252 );setf
253 );multiple-value-bind
254 ;replicates the labels for the begin, middle and end of each
       protein sequence
255 ;come a o processamento de todas as sequencias
256 (loop for fasta in multifasta do
257 (setf sequence (cadr fasta)
258 sequence-name (car fasta)
259 listof-sequence-names (append listof-sequence-names (list
       sequence-name))
260 sequence-size (length sequence)
261 sequence-histogram (dohistogram sequence)
262 searchspace sequence-size
263 milestone (round (* sequence-size 0.01))
264 milestone (if (zerop milestone) 1 milestone)
265 checkmilestone milestone
266 );setf
267 ;come a o processamento para uma sequencia
268 (loop for pos from 0 to (- (length histoalphabet) 1) do
269 ;medidor de progresso
270 ; (when (= (mod pos checkmilestone) 0)
271 ; (setq checkmilestone (+ checkmilestone milestone))
272 ;(print (round (* 100 (/ checkmilestone (- searchspace 1)))))
273 ; );when checkmilestone
274 (setf (aref *printlist* (+ sequence-number 1) (+ pos 1)) (nth pos
        sequence-histogram))
275 );loop
276 ;after the histogram calculation we start to compute the
       physicochemicals properties
277 (setf pos (length histoalphabet))
278 (loop for physico in physicochemicals do
279 (setf (aref *printlist* (+ sequence-number 1) (+ pos 1))
280 (loop for weigth-list in (list physico) sum
281 (loop for product in (mapcar #'* sequence-histogram weigth-list)
       sum product)
282 )
283 )
```

```
(incf pos)
)
;Giving a try: lets compute the number of aminoacids just at the region known to host the signal peptide
(setf sequence-histogram (dohistogramini sequence))
(loop for physico in physicochemicals do
(setf (aref *printlist* (+ sequence-number 1) (+ pos 1))
(loop for weigth-list in (list physico) sum
(loop for product in (mapcar #'* sequence-histogram weigth-list) sum product)
)
)
(incf pos)
)
;Giving another try: lets compute the number of aminoacids just at the end of the protein
(setf sequence-histogram (dohistogramend sequence))
(loop for physico in physicochemicals do
(setf (aref *printlist* (+ sequence-number 1) (+ pos 1))
(loop for weigth-list in (list physico) sum
(loop for product in (mapcar #'* sequence-histogram weigth-list) sum product)
)
)
(incf pos)
)
;Giving another try: lets compute the number of aminoacids just at the middle
(setf sequence-histogram (dohistogrammid sequence))
(loop for physico in physicochemicals do
(setf (aref *printlist* (+ sequence-number 1) (+ pos 1))
(loop for weigth-list in (list physico) sum
(loop for product in (mapcar #'* sequence-histogram weigth-list) sum product)
)
)
(incf pos)
)
(incf sequence-number)
);loop for multifasta
;final pretty print
```

```
;write the name of the sequences in the first line of the *
   printlist*
;In the first line, the first column of *printlist* must be an
   empty string. Starting the names from the second column or x=1
(setf (aref *printlist* 0 0 ) "");empty string
;Inserting the attribute names in first line
(loop for y from 1 to (length histoalphabet) do
(setf (aref *printlist* 0 y )
(symbol-name (nth (- y 1) histoalphabet))))
(loop for y from (length histoalphabet) to (length alphabet) do
(setf (aref *printlist* 0 y )
(nth (- y 1) alphabet)))
;Inserting the name of the sequences in the first column of all
   lines
(loop for y from 1 to multifasta-size do (setf (aref *printlist*
   y 0 ) (nth (- y 1) listof-sequence-names)))
;Once we wrote the first row, now we need to write the data
   concerning each sequence name
(loop for x from 0 to multifasta-size do
(loop for y from 0 to (length alphabet) do
(if (aref *printlist* x y)
(format t "~a~a" (aref *printlist* x y) #\Tab )
(format t "~a~a" 0 #\Tab )
))
(format t "~%" )
)
);let
);if
);defun
```

Listing A.1 – Código fonte em lisp

A.2

Exemplo de formatação das proteínas convertidas para o formato ARFF do WEKA de acordo com a metodologia desse projeto de pesquisa.

@relation localsubcellular

@attribute A numeric
@attribute L numeric
@attribute R numeric

@attribute K numeric
@attribute N numeric
@attribute M numeric
@attribute D numeric
@attribute F numeric
@attribute C numeric
@attribute P numeric
@attribute Q numeric
@attribute S numeric
@attribute E numeric
@attribute T numeric
@attribute G numeric
@attribute W numeric
@attribute H numeric
@attribute Y numeric
@attribute I numeric
@attribute V numeric
@attribute BASIC numeric
@attribute ACID numeric
@attribute POLAR numeric
@attribute NONPOLAR numeric
@attribute CHAM830103 numeric
@attribute CHAM830104 numeric
@attribute CHAM830105 numeric
@attribute FAUJ880111 numeric
@attribute MONM990201 numeric
@attribute MONM990101 numeric
@attribute BEGF750101 numeric
@attribute BROC820102 numeric
@attribute FAUJ880112 numeric
@attribute GEIM800103 numeric
@attribute GEIM800105 numeric
@attribute LEWP710101 numeric
@attribute NAKH900102 numeric
@attribute NAKH900108 numeric
@attribute OOBM850104 numeric
@attribute PALJ810115 numeric
@attribute PONP800106 numeric
@attribute QIAN880116 numeric

@attribute RICJ880107 numeric
@attribute ROBB760111 numeric
@attribute ROSM880103 numeric
@attribute VENT840101 numeric
@attribute AURR980101 numeric
@attribute AURR980105 numeric
@attribute AURR980118 numeric
@attribute ZHOH040103 numeric
@ATTRIBUTE attribute36 {,NEGATIVE,POSITIVE}

@data

28,21,5,17,3,8,16,23,9,13,30,7,4,7,10,15,15,2,3,36,37,33,72,130,275,172,78,37,236.5,260.4,173.62003,
768.80005,33,270.68,280.43002,82.95999,778.8501,-18.73,-445.95,245.05,3213.6396,-21.099998,
306.30002,-241.70001,144.79999,91,277.12003,252.46,270.44003,3593.0999,POSITIVE
44,16,22,24,5,16,36,28,11,21,45,31,11,18,14,23,26,3,12,45,58,60,132,201,457,311,155.5,58,402.80002,
440.20004,286.03003,1097.1,60,480.43,451.21002,136.14001,1267.2098,8.489993,-859.43994,
412.79996,5296.62,-37.91,477,-373.59998,263,144,456.81995,426.71,454.95996,5856,NEGATIVE

A.3

Índice de propensão de aminoácidos disponíveis no repositório AAindex. Os valores representam respectivamente os aminoácidos: A R N D C Q E G H I L K M F P S T W Y V

BASIC: 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
ACID: 0.00 0.00 0.00 1.00 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
POLAR: 0.00 0.00 1.00 0.00 1.00 1.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00 1.00 0.00 1.00 0.00
NONPOLAR: 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00 1.00 0.00 1.00 1.00 1.00 0.00 0.00 1.00 0.00 1.00
CHAM830103: 0.0 1.0 1.0 1.0 1.0 1.0 1.0 0.0 1.0 2.0 1.0 1.0 1.0 1.0 0.0 1.0 2.0 1.0 1.0 2.0
CHAM830104: 0.0 1.0 1.0 1.0 0.0 1.0 1.0 0.0 1.0 1.0 2.0 1.0 1.0 1.0 0.0 0.0 0.0 1.0 1.0 0.0
CHAM830105: 0.0 1.0 0.0 0.0 0.0 1.0 1.0 0.0 1.0 0.0 0.0 1.0 1.0 1.0 0.0 0.0 0.0 1.5 1.0 0.0
FAUJ880111: 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
MONM990201: 0.4 1.5 1.6 1.5 0.7 1.4 1.3 1.1 1.4 0.5 0.3 1.4 0.5 0.3 1.6 0.9 0.7 0.9 0.9 0.4
MONM990101: 0.5 1.7 1.7 1.6 0.6 1.6 1.6 1.3 1.6 0.6 0.4 1.6 0.5 0.4 1.7 0.7 0.4 0.7 0.6 0.5
BEGF750101: 1.00 0.52 0.35 0.44 0.06 0.44 0.73 0.35 0.60 0.73 1.00 0.60 1.00 0.60 0.06

0.35 0.44 0.73 0.44 0.82
BROC820102: 3.9 3.2 -2.8 -2.8 -14.3 1.8 -7.5 -2.3 2.0 11.0 15.0 -2.5 4.1 14.7 5.6 -3.5 1.1 17.8 3.8 2.1
FAUJ880112: 0.00 0.00 0.00 1.00 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
GEIM800103: 1.55 0.20 1.20 1.55 1.44 1.13 1.67 0.59 1.21 1.27 1.25 1.20 1.37 0.40 0.21 1.01 0.55 1.86 1.08 0.64
GEIM800105: 0.84 1.04 0.66 0.59 1.27 1.02 0.57 0.94 0.81 1.29 1.10 0.86 0.88 1.15 0.80 1.05 1.20 1.15 1.39 1.56
LEWP710101 0.22 0.28 0.42 0.73 0.20 0.26 0.08 0.58 0.14 0.22 0.19 0.27 0.38 0.08 0.46 0.55 0.49 0.43 0.46 0.08
NAKH900102: 3.73 3.34 2.33 2.23 2.30 2.36 3.00 3.36 1.55 2.52 3.40 3.36 1.37 1.94 3.18 2.83 2.63 1.15 1.76 2.53
NAKH900108: -0.70 -0.91 1.28 -0.93 -0.41 -0.71 -1.13 -0.12 0.04 1.77 1.02 -0.40 0.86 1.29 -0.42 0.14 -0.13 0.26 1.29 -0.19
OOBM850104: -2.49 2.55 2.27 8.86 -3.13 1.79 4.04 -0.56 4.22 -10.87 -7.16 -9.97 -4.96 -6.64 5.19 -1.60 -4.75 -17.84 9.25 -3.97
PALJ810115 0.91 0.77 1.32 0.90 0.50 1.06 0.53 1.61 1.08 0.36 0.77 1.27 0.76 0.37 1.62 1.34 0.87 1.10 1.24 0.52
PONP800106: 10.67 11.05 10.85 10.21 14.15 11.71 11.71 10.95 12.07 12.95 13.07 9.93 15.00 13.27 10.62 11.18 10.53 11.41 11.52 13.86
QIAN880116: -0.19 0.03 0.02 -0.06 -0.29 0.02 -0.10 0.19 -0.16 -0.08 -0.42 -0.09 -0.38 -0.32 0.05 0.25 0.22 -0.19 0.05 -0.15
RICJ880107: 1.1 1.5 0.0 0.3 1.1 1.3 0.5 0.4 1.5 1.1 2.6 0.8 1.7 1.9 0.1 0.4 0.5 3.1 0.6 1.5
ROBB760111: -3.7 1.0 -0.6 -0.6 4.0 3.4 -4.3 5.9 -0.8 -0.5 -2.8 1.3 -1.6 1.6 -6.0 1.5 1.2 6.5 1.3 -4.6
ROSM880103: 0.4 0.3 0.9 0.8 0.5 0.7 1.3 0.0 1.0 0.4 0.6 0.4 0.3 0.7 0.9 0.4 0.4 0.6 1.2 0.4
VENT840101: 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00 1.00 0.00 0.00 1.00 0.00 0.00 0.00 1.00 1.00 1.00
AURR980101: 0.94 1.15 0.79 1.19 0.60 0.94 1.41 1.18 1.15 1.07 0.95 1.03 0.88 1.06 1.18 0.69 0.87 0.91 1.04 0.90
AURR980105: 0.67 0.76 1.28 1.58 0.37 1.05 0.94 0.98 0.83 0.78 0.79 0.84 0.98 0.96 1.12 1.25 1.41 0.94 0.82 0.67
AURR980118: 0.93 0.96 0.82 1.15 0.67 1.02 1.07 1.08 1.40 1.14 1.16 1.27 1.11 1.05 1.01 0.71 0.84 1.06 1.15 0.74
ZHOH040103: 13.4 8.5 7.6 8.2 22.6 8.5 7.3 7.0 11.3 20.3 20.8 6.1 15.7 23.9 9.9 8.2 10.3 24.5 19.5 19.5