UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO" - UNESP

FACULDADE DE ENGENHARIA

CÂMPUS DE ILHA SOLTEIRA

ANTHONY FERREIRA LA MARCA

Uma Ferramenta Para Predições de Epítopos Lineares de Células B Baseada Em Uma Rede Neural da Teoria da Ressonância Adaptativa

Ilha Solteira
2022

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Uma Ferramenta Para Predições de Epítopos Lineares de Células B Baseada Em Uma Rede Neural da Teoria da Ressonância Adaptativa

Anthony Ferreira La Marca

Tese de Doutorado apresentada à Faculdade de Engenharia de Ilha Solteira – UNESP como parte dos requisitos para obtenção do título de Doutor em Engenharia Elétrica.

Área de Conhecimento: Automação

Prof. Dr. Carlos Roberto Minussi
Orientador

Ilha Solteira
2022

FICHA CATALOGRÁFICA
Desenvolvido pelo Serviço Técnico de Biblioteca e Documentação

M313f

Marca, Anthony Ferreira La.
Uma ferramenta para predições de epítopos lineares de células B baseada em uma rede neural da teoria da ressonância adaptativa / Anthony Ferreira La Marca. -- Ilha Solteira: [s.n.], 2022
82 f. : il.

Tese (doutorado) - Universidade Estadual Paulista. Faculdade de Engenharia de Ilha Solteira. Área de conhecimento: Automação, 2022

Orientador: Carlos Roberto Minussi
Coorientador: Robson da Silva Lopes
Inclui bibliografia

1. Mapeamento de epítopos. 2. Diagnóstico. 3. Predição *in* sílico. 4. Epítopos lineares de células B. 5. Artmap-fuzzy.

João Josué Barbosa
Serviço Técnico de Biblioteca e Documentação
Diretor Técnico
CRB 8-5642

unesp

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Ilha Solteira

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: Uma Ferramenta Para Predição de Epítopos Lineares de Células B Baseada Em Uma Rede Neural da Teoria da Ressonância Adaptativa

AUTOR: ANTHONY FERREIRA LA MARCA
ORIENTADOR: CARLOS ROBERTO MINUSSI

Aprovado como parte das exigências para obtenção do Título de Doutor em ENGENHARIA ELÉTRICA, área: Automação pela Comissão Examinadora:

Prof. Dr. CARLOS ROBERTO MINUSSI (Participaçao Virtual)
Departamento de Engenharia Elétrica / Faculdade de Engenharia de Ilha Solteira - UNESP

Profª. Drª. ANNA DIVA PLASENCIA LOTUFO (Participaçao Virtual)
Departamento de Engenharia Elétrica / Faculdade de Engenharia de Ilha Solteira - UNESP

Prof. Dr. ROBSON DA SILVA LOPES (Participaçao Virtual)
Bioinformática / Universidade Federal de Mato Grosso

Prof. Dr. KENJI NOSE FILHO (Participaçao Virtual)
Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas / Universidade Federal do ABC - UFABC

Dr. JOÃO LUÍS REIS CUNHA (Participaçao Virtual)
Bioinformática / University of York

Ilha Solteira, 11 de fevereiro de 2022

DEDICATÓRIA

A DEUS

À minha família

AGRADECIMENTOS

Primeiramente agradeço a Deus por me fortalecer nos momentos de dificuldade, estar sempre ao meu lado guiando meus passos, e me presentear com pessoas tão especiais que me ensinam a cada dia a me tornar uma pessoa melhor.

A meus pais Mauri La Marca e Vanda Ferreira, bem como meu tio Wilson La Marca e minha vó Maria La Marca pela dedicação, apoio, confiança, carinho, amizade e amor em todos os momentos.

A minha esposa Mayara Cristina Florêncio La Marca pelo carinho, amor e compreensão durante toda essa etapa da minha vida.

Aos meus colegas de laboratório, pelos conselhos e companheirismo.

A todos meus companheiros de profissão do curso de Ciência de Computação da Universidade Federal do Mato Grosso, campus Araguaia, que foram compreensivos e permitiram o meu afastamento.

Ao professor Dr. Robson da Silva Lopes e à professora Dr. Daniella Castanheira Bartholomeu pela contribuição ao trabalho de diversas maneiras.

As professoras Anna Diva e Mara Lopes que contribuíram indiretamente com o desenvolvimento desse trabalho, além da amizade e da compreensão em alguns momentos dessa jornada.

E principalmente ao professor Dr. Carlos Roberto Minussi pela orientação, amizade, apoio, paciência e profissionalismo no desenvolver desse trabalho, além da confiança em orientar alguém que veio de outra área de conhecimento.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

ART	*Adaptive Resonance Theory*
AUC	*Area Under the Curve*
BCR	Receptor de Linfócito B
BepFAMN	*B Epitop Prediction Fuzzy ArtMap artificial Neural network*
DNN	Redes Neurais Profundas
ELISA	*Enzyme Linked Immunosorbent Assay*
ELISpot	*Enzyme Linked immunospot*
HIV	*Human Immunodeficiency Virus*
IEDB	Immune Epitope Databases
ITAMs	Motivos de Ativação Baseados em Imunorreceptores
ITIM	Inibição Baseada em Imunorreceptores de Tirosina
Ig	Imunoglobulina
LSTM	*Long Short-Term Memory*
MHC	*Major Histocompatibility Complex*
MCC	Coeficiente de Correlação Matthews
PAMPs	Padrões Moleculares Associados a Patógenos
PSSM	*Position-specific scoring matrix*
RRP	Receptores de Reconhecimento de Padrões
RF	*Randon Forest*
RNA	Rede Neural Artificial
RNAs	Redes Neurais Artificiais
RNN	Redes Neurais Recorrentes
ROC	*Receiver Operating Characteristic Curve*
SVM	*Support Vector Machine*

LISTA DE SÍMBOLOS

F_0	Camada de entrada
F_1	Camada de comparação
F_2	Camada de reconhecimento
$\boldsymbol{a}$	Vetor de entrada
$\boldsymbol{I}$	Vetor de atividade da camada F_0
x	Vetor de atividade da camada F_1
y	Vetor de atividade da camada F_2
ρ	Parâmetro de vigilância
β	Taxa de Treinamento
α	Parâmetro de escolha
M	Número de neurônios da camada F_0
N	Número de neurônios da camada F_2
T_j	Função de escolha
J	Categoria ativa da camada F_2
$\mathbf{w_{ij}}$	Matriz de pesos
$\boldsymbol{b}$	Vetor de entrada do módulo ART_b
F_0^a	Camada de entrada do módulo ART_a
F_1^a	Camada de comparação do módulo ART_a
F_2^a	Camada de reconhecimento do módulo ART_a
F_0^b	Camada de entrada do módulo ART_b
F_1^b	Camada de comparação do módulo ART_b
F_2^b	Camada de reconhecimento do módulo ART_b
F_b^a	Camada Inter-ART
$\boldsymbol{I_a}$	Vetor de atividade da camada F_0 do módulo ART_a
$\boldsymbol{I_b}$	Vetor de atividade da camada F_0 do módulo ART_a
$\boldsymbol{x_a}$	Vetor de atividade da camada F_1 do módulo ART_b
$\boldsymbol{x_b}$	Vetor de atividade da camada F_1 do módulo ART_b
$\boldsymbol{x_{ab}}$	Vetor de atividade do módulo Inter-ART
$\boldsymbol{y_b}$	Vetor de atividade da camada F_2 do módulo ART_b
ρ_a	Parâmetro de vigilância do módulo ART_a

$\bar{\rho}_{\mathrm{a}}$	Parâmetro de vigilância base do módulo $\mathrm{ART_a}$
ϵ	Valor de incremento e decremento do parâmetro ρ_{a}
M_a	Número de neurônios da camada F_0 do módulo $\mathrm{ART_a}$
M_b	Número de neurônios da camada F_0 do módulo $\mathrm{ART_a}$
N_a	Número de neurônios da camada F_2 do módulo $\mathrm{ART_b}$
N_b	Número de neurônios da camada F_2 do módulo $\mathrm{ART_b}$
K	Categoria ativa da camada F_2 do módulo $\mathrm{ART_b}$
$\boldsymbol{w_j^a}$	Vetor de pesos do módulo $\mathrm{ART_a}$
$\boldsymbol{w_k^b}$	Vetor de pesos do módulo $\mathrm{ART_b}$
$\boldsymbol{w_{jk}^{ab}}$	Matriz de pesos do módulo Inter-ART

RESUMO

O sistema público de saúde é extremamente dependente do uso de vacinas para imunizar a população de uma série de doenças infecciosas e perigosas, evitando que o sistema entre em colapso e que milhões de pessoas morram todo ano. No entanto, para desenvolvê-las e monitorar de forma efetiva essas doenças é necessário utilizar métodos de diagnóstico precisos, capazes de identificar regiões altamente imunogênicas dentro de uma determinada proteína patogênica. Os métodos experimentais existentes têm custos elevados, são demorados e exigem um árduo trabalho laboratorial, pois requerem a triagem de um grande número de potenciais epítopos candidatos, tornando os métodos extremamente laboriosos, especialmente para a aplicação em microrganismos maiores. Nas últimas décadas, os pesquisadores desenvolveram métodos de predição *in sílico*, baseados em aprendizagem de máquina, para identificar esses marcadores, de maneira a reduzir drasticamente a lista de potenciais epítopos candidatos para os testes experimentais, e, consequentemente, diminuir a laboriosa tarefa associada ao seu mapeamento. Apesar dos esforços da última década e da quantidade de dados disponíveis em grandes bases de dados públicas, as ferramentas e métodos desenvolvidos, com o propósito de identificar esses marcadores, ainda apresentam baixa acurácia, aprendizado lento e não utilizam técnicas de aprendizado on-line. Desta forma, a proposta deste trabalho é desenvolver uma ferramenta que utilize uma abordagem inédita, atentando ao treinamento on-line e na melhora da acurácia na identificação de epítopos lineares de células B. Para isso, a ferramenta nomeada BepFAMN (*B Epitop Prediction Fuzzy ArtMap Artificial Neural*) network, utiliza a Rede Neural Artificial (RNA) ARTMAP-FUZZY, treinada com epítopos anotados de sequências de aminoácidos de proteína, disponíveis no banco de dados do IEDB. Essa base de dados foi particionada utilizando a técnica de validação cruzada quíntupla e operada para treinamento e validação, sendo que os dados, antes de serem apresentados à RNA, foram pré-processados utilizando a escala de propensão de aminoácidos e sua proporção em epítopos positivos e negativos. Para os testes foi utilizado a base de dados do BepiPred-2.0, como uma base independente. Em ambos, validação e teste, os resultados foram promissores, alcançando área sob a curva (AUC) ROC de aproximadamente 0,9289 e 0,7831, respectivamente. Os valores alcançados, principalmente o de teste, demonstram que os melhores resultados, até então alcançados pela ferramenta EpiDope (0,605), foram superados. Este fato, contribui com uma redução considerável do número de potenciais epítopos lineares a serem validados experimentalmente, reduzindo o tempo laboratorial e acelerando o desenvolvimento de testes de diagnósticos, vacinas e abordagens imunoterapêuticas.

Palavras-chave: mapeamento de epítopo; diagnóstico; predição in sílico; epítopos lineares de células B; ARTMAP-FUZZY

ABSTRACT

The public health system is extremely dependent on the use of vaccines to immunize the population from a range of infectious and dangerous diseases, preventing the system from collapsing and millions of people dying every year. However, to effectively develop and monitor these diseases, it is necessary to use precise diagnostic methods capable of identifying highly immunogenic regions within a particular pathogenic protein. Existing experimental methods have high costs, are time-consuming and require hard laboratory work, as they require the screening of large number of potential candidate epitopes, making the methods extremely laborious, especially for application on larger microorganism. In the last decades, researchers have developed in silico prediction methods, machine learning based, to identify these markers, in order to drastically reduce the list of potential candidate epitopes for the experimental tests, and, consequently, decrease the laborious task associated with their mapping. Despite the efforts of the last decade and the amount of data available in large public databases, the tools and methods developed, with the purpose of identifying these markers, still have low accuracy. Thus, the purpose of this work is to develop a tool that uses an unprecedented approach, focusing on online training and improving the accuracy of identifying linear B-cell epitopes. For this, the tool called BepFAMN, uses the Fuzzy-ARTMAP Artificial Neural Network (ANN) trained considering annotated epitopes from sequences of protein amino acids, available from the IEDB database (Immune Epitope Databases). This database was partitioned using the five-fold cross-validation technique and operated for training and testing, and the data, before being presented to the ANN, were pre-processed using the amino acid propensity scale and its proportion in epitopes positive and negative. For the tests, the BepiPred-2.0 database was used, as an independent database. In both, validation and test, the results were promising, reaching area under the curve (AUC) ROC of approximately 0.9289 and 0.7831, respectively. The achieved values, especially the test one, demonstrates that the best results, hitherto achieved by the EpiDope tool (0.605), were surpassed. This fact contributes to a considerable reduction in the number of potential linear epitopes to be experimentally validated, reducing laboratory time and accelerating the development of diagnosis tests, vaccines and immunotherapeutic approaches.

Keywords: epitope mapping; diagnosis; in silico prediction; linear B-cell epitopes; artificial neural network; fuzzy-ARTMAP.

1 – INTRODUÇÃO

Para evitar que o sistema de saúde entre em colapso e que milhões de pessoas morram anualmente, é importante desenvolver testes de diagnósticos precisos, que ajudem no controle e no monitoramento, bem como vacinas eficientes que imunizem as pessoas contra doenças perigosas e altamente infecciosas.

Para desenvolver vacinas e diagnósticos baseados em peptídeos e oferecer maior segurança, potência e eficácia (EL-MANZALAWY; HONAVAR, 2010), primeiramente, é necessário identificar as regiões altamente imunogênicas dentro de uma determinada proteína de organismos patogênicos. Essas regiões que representam a interface entre o agente patogênico e a resposta imune, são conhecidas como epítopos de células B e T, responsáveis por induzirem uma resposta imune (KRINGELUM *et al*., 2013).

Os epítopos de células B são classificados como lineares ou conformacionais. Nos lineares o epítopo possui aminoácidos sequenciais (contínuos), enquanto o epítopo conformacional é composto por aminoácidos que não sejam necessariamente contínuos na sequência primária, são colocados próximos por meio do dobramento da proteína (VAN REGENMORTEL, 2009). Embora grande parte dos epítopos sejam conformacionais, sua identificação requer informações baseadas na estrutura tridimensional que ainda são muito limitadas comparadas ao número de sequências de proteínas atualmente disponíveis em bancos de dados públicos (KRINGELUM *et al*., 2013). Portanto, este estudo concentra-se na previsão de epítopos lineares de células B.

Os métodos convencionais e mais confiáveis utilizados para o mapeamento desses epítopos, como cristalografia (RUX; BURNETT, 2000) e técnicas de Ressonância Magnética Nuclear (MAYER; MEYER, 2001), são custosos e exigem um extenso trabalho laboratorial. Logo, os métodos de identificação *in sílicos* têm sido amplamente utilizados para o desenvolvimento de modelos computacionais que possam prever com mais eficácia a presença e a localização de epítopos lineares de células B, a partir de uma sequência de aminoácidos de um organismo patogênico (SUN *et al*., 2019).

Devido a simplicidade na geração de *embeddings*, o estudo da predição de epítopos lineares de células B vem ocorrendo há vários anos e já conta com muitas ferramentas disponíveis na literatura, nas quais utilizam técnicas de aprendizado de máquina em suas predições, tais como: ABCPred (SAHA; RAGHAVA, 2006), BCPred (EL-MANZALAWY; HONAVAR, 2008), SVMTrip (YAO *et al*., 2012), BepiPred-2.0 (JESPERSEN *et al*., 2017) e

EpiDope (COLLATZ *et al*., 2020). No entanto, os desempenhos ainda são bastante limitados, o que não garante a identificação altamente precisa de epítopos lineares de células B.

É importante salientar que vários fatores podem influenciar esta baixa acurácia, sendo que um deles está relacionado diretamente com a quantidade e a qualidade dos dados disponíveis para o treinamento. Assim, com a atualização constante de uma das principais bases de dados pública, a *Immune Epitope Databases* (IEDB) (VITA *et al*., 2018), unido a novas tendências de extração de características dos dados e de novas técnicas de Inteligência Artificial, espera-se que gradualmente a acurácia na identificação de epítopos lineares de células B, atinja patamares de excelência.

Diante deste contexto, esta pesquisa tem por objetivo apresentar uma nova ferramenta computacional, chamada BepFAMN, que identifica e localiza epítopos lineares de células B em sequências de aminoácidos. A ferramenta utilizou uma RNA da família ART (*Adaptive Resonance Theory*) que permite o treinamento on-line, além de prover grande facilidade na aprendizagem de novos padrões, sem perder a memória dos padrões já aprendidos (plasticidade e estabilidade). Sobre a base de dados retirada do IEDB (em 2021) foi aplicado o software BLAST[1] (CAMACHO *et al*., 2013), para redução de redundância, 14 escalas de propensão de aminoácidos foram extraídas, propostas por (LIN, 2013) e foi calculado a proporção dos resíduos presentes em epítopos positivos e epítopos negativos. Todo esse processo de geração dos *embeddings* foi feito por meio de uma janela deslizante, de tamanho pré-fixado.

Para treinar e validar a RNA ARTMAP-FUZZY (o algoritmo usado), foi utilizado a técnica de validação cruzada quíntupla. As métricas empregadas para avaliar a qualidade da ferramenta BepFAMN, foram: Sensibilidade, Especificidade, Acurácia, Coeficiente de Correlação Matthews[2] (MCC) e Área sob a Curva (AUC).

Para os testes foi utilizado a base de dados do BepiPred-2.0 (JESPERSEN *et al*., 2017) e os modelos de conhecimento gerados pelo conjunto de dados de treinamento. Inicialmente, foram aplicados os pré-processamentos descritos no trabalho de (COLLATZ et. al., 2020), a fim de deixar o conjunto de dados idêntico ao utilizado pelo autor. Após esse processo, foram gerados os embeddings (as entradas da RNA), para posteriormente mensurar e comparar os resultados. Nesta etapa de teste foram utilizadas todas as métricas supracitadas, com a adição da AUC da curva *Precision-Recall*, pois, segundo (DAVIS; GOADRICH, 2006), em situações

[1] Segundo (ALTSCHUL *et al*., 1990), BLAST é um software que encontra regiões de similaridade local entre sequências de nucleotídeos ou proteínas.

[2] Segundo (MATTHEWS, 1975), MCC é um coeficiente utilizado em *machine learning* para avaliar a qualidade de classificação binária. Ele retorna valores entre -1 e 1, onde -1 representa discordância total entre previsão e observação, 0 representa uma previsão próxima de uma aleatória e 1 representa uma previsão perfeita.

em que as classes estão desbalanceadas (quantidade de epítopos negativos muito superior a quantidade de epítopos positivos, nesse estudo), essa medida fornece uma visão mais real do desempenho do algoritmo.

Uma AUC da curva ROC de aproximadamente 0,7831 e uma AUC da curva *Precision-Recall* de aproximadamente 0,8343, retrata que a ferramenta BepFAMN superou os métodos já desenvolvidos, inclusive a ferramenta EpiDope (considerada uma das melhores ferramentas), que tinha atingido valores de 0,605 e 0.660, respectivamente.

O trabalho está composto da seguinte maneira: a seção 2 apresenta os principais conceitos teóricos que norteiam o presente estudo; a seção 3 apresenta os trabalhos correlatos, a seção 4 apresenta a RNA ARTMAP-FUZZY, a seção 5 apresenta os materiais e métodos utilizados para o desenvolvimento da ferramenta, bem como os bancos de dados, os pré-processamentos, a estratégia de predição, a arquitetura da RNA usada, o seu treinamento e diagnóstico e as métricas utilizadas; a seção 6 apresenta os resultados do treinamento, da validação e dos testes; a seção 7 aborda as discussões; e por fim, a seção 8 destaca as conclusões e direcionamentos para trabalhos futuros.

1.2 – Objetivo Geral

Desenvolver uma ferramenta computacional capaz de identificar e localizar epítopos lineares de células B em sequências de proteína, que supere os resultados das principais ferramentas disponíveis na literatura e

1.3 – Objetivos Específicos

- Criar uma base de dados local de epítopos lineares de células B de bactérias, de vírus e de protozoários, sendo todos identificados experimentalmente e disponíveis no banco de dados IEDB;
- Criar uma cópia da base de dados local e dividi-la em três subconjuntos, sendo cada um formado exclusivamente por dados de um táxon;
- Implementar a RNA ARTMAP-FUZZY e aplicá-la na base de dados local e individualmente em cada subconjunto de dados;
- Investigar e comparar as predições dos subconjuntos de dados e da base de dados local;

- Reconstruir a base de dados do trabalho (COLLATZ *et al*., 2020), incluindo todos os seus pré-processamentos, para criar um conjunto de dados análogo e utilizá-lo para validação;
- Aplicar o diagnóstico da RNA ARTMAP-FUZZY no conjunto de dados de validação, utilizando o modelo de conhecimento gerado pela base de dados local;
- Comparar e analisar os resultados das predições do conjunto de dados de validação com os resultados das principais ferramentas de predição de epítopos lineares de células B, disponíveis na literatura.

2 – REVISÃO DA LITERATURA

2.1 - Imunidade Inata e Adaptativa

Historicamente, o termo imunidade significa proteção contra doenças infecciosas, sendo que nos seres humanos, sistema imune é composto por células e moléculas responsáveis pela imunidade. Sua resposta coordenada e coletiva à entrada de substâncias estranhas em nosso corpo é chamada de reposta imune. No entanto, mesmo substâncias não infecciosas, como lesões causadas pelos mecanismos protetores na eliminação de substâncias estranhas, às vezes, as próprias moléculas do indivíduo (doenças autoimunes) podem induzir às repostas imunes. Portanto, de uma forma mais ampla, resposta imune é uma reação aos componentes de microrganismos, como proteínas e polissacarídeos, além de agentes químicos que são reconhecidos como estranhos (ABBAS; LICHTMAN, 2019).

A imunidade inata atua em conjunto com a imunidade adaptativa e caracteriza-se pela rápida resposta à agressão, independentemente do estímulo prévio, sendo a primeira linha de defesa do organismo. Ela consiste em mecanismos de defesas celulares e bioquímicos, como barreiras físicas, químicas e biológicas, componentes celulares e moléculas solúveis, que estão ativos mesmo sem haver a infecção. Após a infecção, não se alteram quantitativa, ou qualitativamente (MEDZHITOV; JANEWAY, 2000).

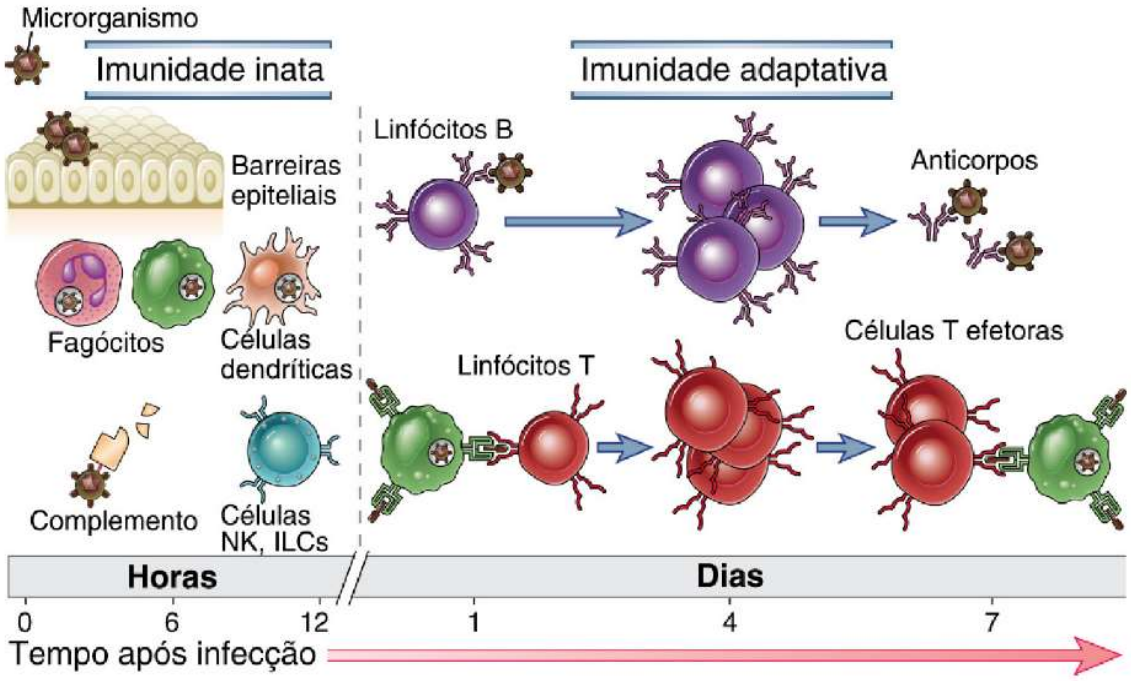
Macrófagos, neutrófilos, células dendríticas e células *Natural Killer* são as principais células efetoras da imunidade inata, sendo que os seus principais mecanismos são compostos pela fagocitose, ativação de proteínas do sistema complemento, citocinas, quimiocinas e liberação de mediadores inflamatórios. Esses mecanismos são ativados por estímulos peculiares, representados por estruturas moleculares, denominados Padrões Moleculares Associados a Patógenos (PAMPs). Tais padrões ativam a reposta imune inata, por meio do contato com diferentes receptores, os quais são conhecidos como Receptores de Reconhecimento de Padrões (RRP). Essa interação reconhece apenas os padrões moleculares já projetados no código genético, não havendo diversidade nem capacidade adaptativa para gerar novos receptores que reconheçam novos padrões (ABBAS; LICHTMAN, 2019).

Com a evolução dos microrganismos patogênicos, muitos começaram a resistir à imunidade inata, fato que desencadeou a necessidade de um sistema imune mais potente, diversificado e adaptativo para eliminá-los.

A imunidade adaptativa é ativada pela exposição a antígenos, que são substâncias estranhas reconhecidas pelos linfócitos ou moléculas solúveis secretadas por ele, como

anticorpos e citocinas. O sistema imune adaptativo é capaz de reconhecer e reagir a inúmeros antígenos. Além disso, à medida que essa exposição se repete, as células de memória imunológica aumentam significativamente a capacidade defensiva para um determinado microrganismo (TRAVIS, 2009). A Figura 1 apresenta os mecanismos de defesas iniciais da imunidade inata e as respostas imunes adaptativas desenvolvidas posteriormente.

Figura 1 - Imunidade Inata e Adaptativa



Fonte: (ABBAS; LICHTMAN, 2019)

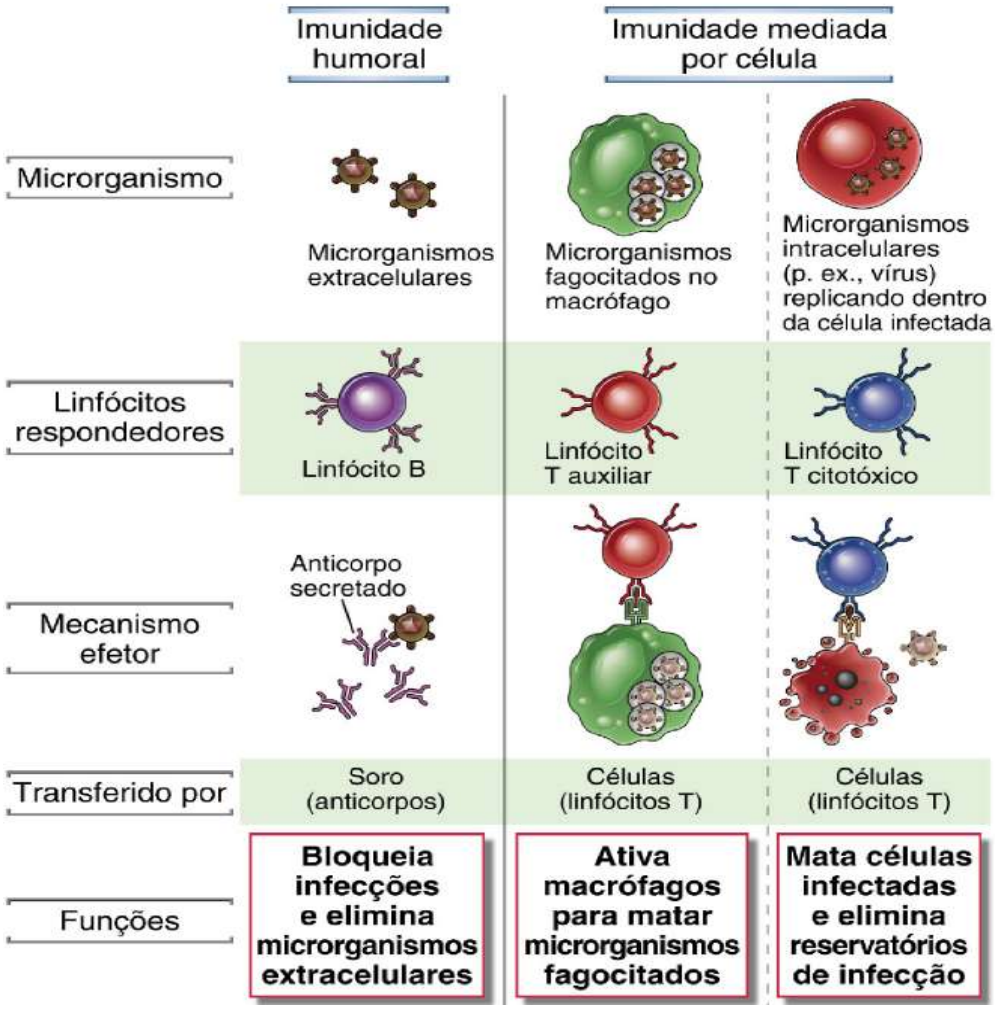
De acordo com a Figura 1, ao primeiro contato com o microrganismo a imunidade inata ativa os seus mecanismos efetores, como fagócitos e células dendríticas, tentando eliminá-lo. Ao decorrer da infecção, caso esse primeiro enfrentamento não seja efetivo, o antígeno estimula os linfócitos que ativam os seus mecanismos efetores para gerar anticorpos e células T efetoras que atacam e tentam erradicar os invasores.

As respostas imunes adaptativas atuam de diferentes formas e utilizam diferentes componentes do sistema imune para eliminar vários tipos de microrganismos. Essa maneira de atuação é classificada como imunidade humoral e imunidade celular (SILVERSTEIN, 2003).

A imunidade humoral utiliza anticorpos, produzidos e secretados pelos linfócitos B (célula B), para reconhecer, neutralizar e eliminar os microrganismos extracelulares e seus produtos. Os anticorpos podem ativar diferentes mecanismos efetores para combater os microrganismos, como, por exemplo, promover a sua ingestão pelas células do hospedeiro

(fagocitose). A imunidade celular é mediada pelos linfócitos T (células T) e seus produtos, que reconhecem e eliminam os microrganismos que sobrevivem e se proliferam dentro das células do hospedeiro (intracelular), como vírus e algumas pequenas bactérias. Os linfócitos T auxiliares contribuem para a destruição desses microrganismos, enquanto os linfócitos T citotóxicos eliminam a célula infectada. Os linfócitos T auxiliares também auxiliam os linfócitos B na eliminação de microrganismos extracelulares, convocando leucócitos que os destroem (SILVERSTEIN, 2003). A Figura 2 apresenta os tipos de imunidade adaptativas brevemente discutidos.

Figura 2 - Tipos de Imunidades Adaptativas



Fonte: (ABBAS; LICHTMAN, 2019)

2.2 – Linfócitos de Células B e Produção de Anticorpos

Os linfócitos são as únicas células do corpo humano que manifestam receptores de antígenos clonalmente distribuídos para atuarem em cada determinante antigênico de forma específica. Esse fato, colabora com a existência de milhões de clones espalhados pelo corpo, permitindo que o sistema imune tenha um repertório extremamente diverso de receptores que reconheçam e respondam a milhões de antígenos diferentes (ABBAS; LICHTMAN, 2019).

Os linfócitos B, as células que produzem anticorpos, possuem características fenotípicas e funcionais diferentes, e, portanto, são subdivididos em três principais grupos: as células B foliculares, as células B da zona marginal e as células B-1. As células B foliculares são as que possuem maior diversidade de anticorpos clonalmente distribuídos e contribuem com receptores de antígenos e mecanismos efetores para a imunidade humoral adaptativa. Já os outros dois grupos, geram anticorpos com diversidade bastante limitada (LITMAN *et al*., 2010).

Os linfócitos B surgem a partir de células tronco na medula óssea (células imaturas) e passam por complexos processos de maturação, nas quais desenvolvem os seus receptores de antígenos e adquirem as suas características fenotípicas e funcionais (células maduras). Todo esse processo de maturação das células B ocorre na medula óssea, diferentemente das células T, que são geradas pela medula óssea, mas migram e amadurecem no Timo[3] (ABBAS; LICHTMAN, 2019).

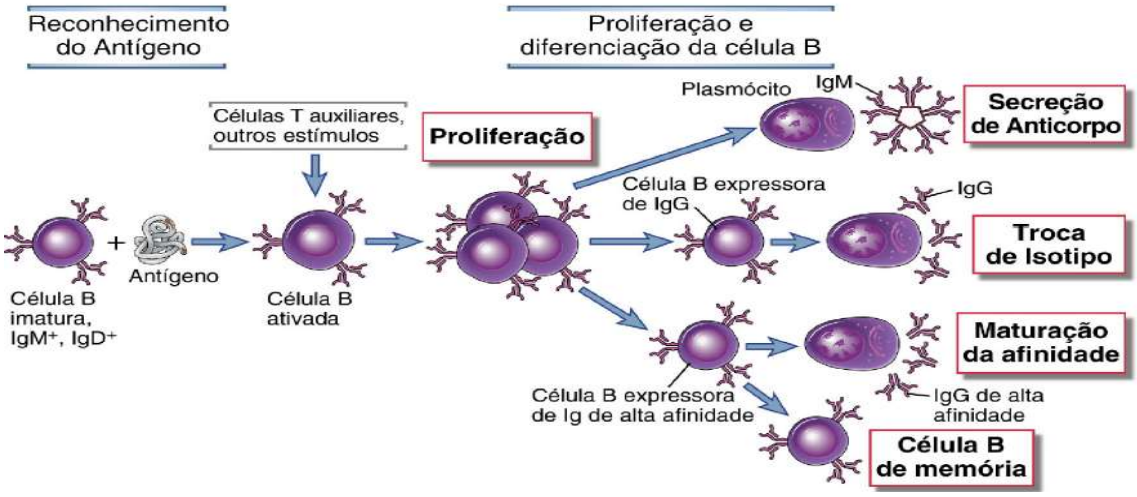
Após os antígenos interagirem com as imunoglobulinas (anticorpos) tipo IgM e IgD das membranas dos linfócitos B virgem, os linfócitos se proliferam e se diferenciam em células efetores e de memória. Devido a rápida replicação dos microrganismos, em sua proliferação, os linfócitos B imaturos transmitem a sua especificidade de reatividade aos seus clones, processo denominado expansão clonal. Em paralelo, os linfócitos se diferem em células efetores e de memória, cujas respectivas funções são eliminar o antígeno e fornecer repostas potentes e rápidas para subsequentes infecções. Os linfócitos B efetores são plasmócitos que secretam anticorpos. Já os linfócitos T efetores, são células que secretam citocinas e células T citotóxicas (ABBAS; LICHTMAN, 2019).

Durante a resposta imune humoral, alguns linfócitos B ativos começam a produzir outros tipos de anticorpos, processo denominado troca de isotipo de cadeia pesada. Além disso, os anticorpos que se ligam com maior afinidade aos antígenos passam a dominar

[3] Segundo (LIMA, SAMPAIO, 2007), o timo é uma glândula linfoide primária especializada do sistema imunológico.

progressivamente a resposta imune (maturação de afinidade) (CERUTTI, 2008). A Figura 3 apresenta as fases da resposta imune humoral.

Figura 3 - Fases da Resposta Imune Humoral



Fonte: (ABBAS; LICHTMAN, 2019)

Segundo a Figura 3, com a ativação das células B pela interação com um antígeno específico, que ocorre por meio de receptores Ig da superfície das células, é iniciada a proliferação e a diferenciação em clones específicos. Os clones descendentes podem se diferenciar em plasmócitos que produzem IgM (o primeiro anticorpo a ser produzido quando há uma infecção) ou outros isotipos de Ig, como o IgG (anticorpo produzido posteriormente, porém mais específico e eficiente à infecção), podendo ainda sofrer maturação de afinidade ou permanecer como células de memória.

Quando ocorre a produção suficiente de anticorpos e ocorre a formação de complexos chamado antígeno-anticorpo, ativa-se uma sinalização inibitória em cascata, por meio de receptores inibitórios, como o FcyRIIB, que finaliza a ativação das células B (ABBAS; LICHTMAN, 2019).

2.3 – Epítopos de Células b

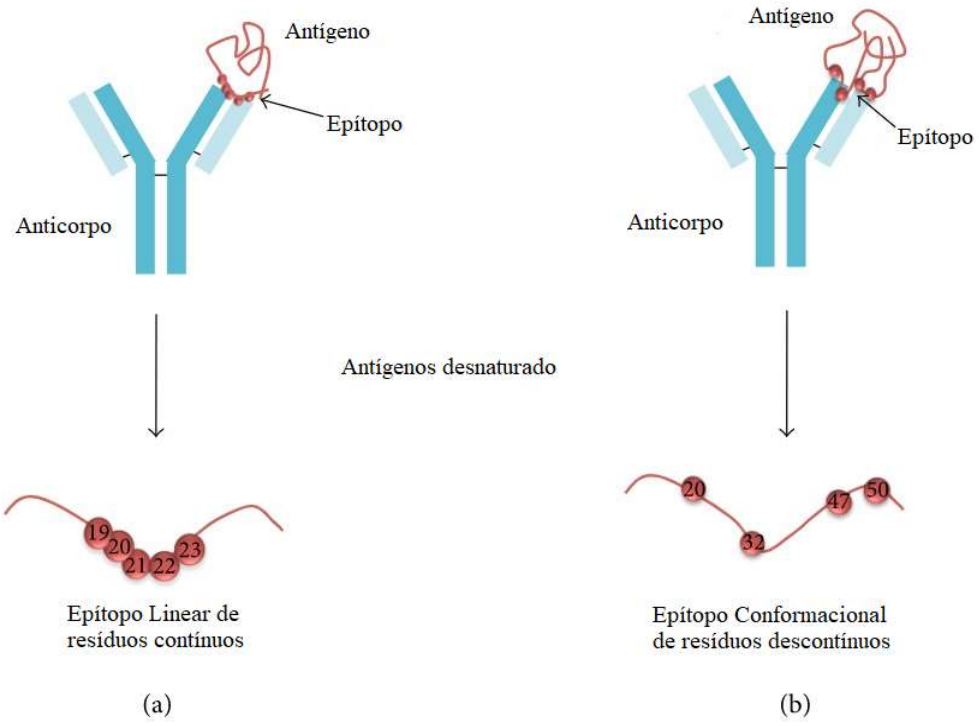
Os linfócitos B e T pertencem a um grupo específico de células, responsáveis por fornecerem as respostas imunes adaptativas do organismo a diversos tipos de antígenos. Em ambos, a resposta imune é estimulada pelo reconhecimento de uma parte específica do antígeno, conhecida como epítopo (ABBAS; LICHTMAN, 2019). Epítopos são regiões específicas do

antígeno reconhecidas pelo sistema imunológico, que se ligam a receptores celulares ou a anticorpos, capazes de despertar uma resposta imune (KINDT *et al*. 2008).

Os epítopos de células B são bastante diversos, podendo ser formados por proteínas, lipídios e carboidratos. Os derivados de proteína são geralmente formados por aminoácidos hidrofílicos, capazes de ser reconhecidos tanto por um anticorpo quanto por um receptor de linfócito B (BCR) (ABBAS; LICHTMAN, 2019). Em razão de estes antígenos estarem livres em solução, seus epítopos tendem a estar em locais altamente acessíveis, geralmente expostos na superfície do antígeno (KINDT *et al*. 2008).

De acordo com o seu arranjo, os epítopos de células B podem ser contínuos (lineares) ou descontínuos (conformacionais), conforme ilustra a Figura 4. Os epítopos contínuos são compostos de aminoácidos sequenciais que se ligam ao anticorpo e, portanto, sua interação é baseada na estrutura primária da proteína, conforme Figura 4 (a). Os conformacionais são formados quando os aminoácidos não contínuos são espacialmente arranjados juntos durante o enovelamento das proteínas, conforme Figura 4 (b). É importante notar que os epítopos descontínuos também podem conter alguns pequenos segmentos contínuos (ABBAS; LICHTMAN, 2019).

Figura 4 - Representação de Epítopo Linear e Conformacional de Célula B. Os lineares (a) são formados a partir de resíduos sequenciais, enquanto os conformacionais (b) contêm resíduos descontínuos ao longo da sequência.



Fonte: (SANCHEZ-TRINCADO; GOMEZ-PEROSANZ, 2017) (Adaptado)

2.4 – Mapeamento de Epítopos de Células B

As tecnologias experimentais têm permitido a identificação de epítopos importantes para uma resposta imune efetiva, denominada mapeamento. O conhecimento dessas regiões é valioso para a compreensão das bases moleculares da imunogenicidade, além de poder utilizá-la para gerar novos testes sorológicos de diagnósticos, imunoterapia e vacinas.

O mapeamento de epítopos de células B em laboratórios pode ser feito por métodos estruturais ou funcionais. Os estruturais incluem a cristalografia, ressonância magnética e microscopia de elétrons (KINDT *et al*. 2008). Os funcionais utilizam técnicas de ressonância de plasma de superfície, espectrometria de massa, varredura de peptídeos baseada em vetor, imunoensaios, como ELISA (do inglês *Enzyme Linked Immunosorbent Assay*) e ELISpot (do inglês *Enzyme Linked immunospot*), Western blot, dentre outros (PONOMARENKO; REGENMORTEL, 2009).

A identificação de muitos epítopos, que tem ocorrido nos últimos anos, deve-se ao surgimento dessas novas abordagens, as várias formas existentes de ensaios de validação e do crescente interesse dos pesquisadores em identificar novos antígenos. Com o propósito de unificar todas estas informações em um único repositório e facilitar pesquisas futuras, em 2005, foi criado um banco de dados chamado IEDB (do inglês, *Immune Epitope Databases*), responsável por armazenar todas essas informações sobre os epítopos de células B e T (VITA *et al*.2010).

Com o desdobramento de novas pesquisas, novos epítopos foram identificados e validados experimentalmente. A fim de manter o repertório do IEDB sempre atualizado, foi desenvolvido algoritmos baseados em máquinas de vetores de suporte[4] (SVM) e redes bayesianas[5], responsáveis por localizarem essas novas publicações científicas no PubMed[6] e atualizarem o seu banco de dados. Ademais, os seus metadados também são armazenados, permitindo realizar buscas mais refinadas sobre o epítopo, tais como: consultar por sua estrutura, por código do organismo ou do antígeno, por restrição de MHC[7] (do inglês, Major Histocompatibility Complex), por tipo de ensaio ou organismo hospedeiro e por doença (ABBAS; LICHTMAN, 2019; VITA *et al*.2010).

[4] Segundo (LORENA, CARVALHO, 2007), SVM é um método de aprendizado de máquina que analisa os dados e reconhece padrões.
[5] Segundo (KORB, NICHOLSON, 2003), redes bayesianas são grafos que representam relações de probabilidade condicional, ou seja, são modelos gráficos para raciocínio baseados em incertezas.
[6] Segundo (NCBI, 2022), PubMed é um motor de busca de livre acesso à base de dados MEDLINE de citações e resumos de artigos de investigações em biomedicina.
[7] Segundo (ABBAS, LICHTMAN, 2019), MHC codifica um grupo de antígenos ou proteínas encontrado na superfície das células.

Os agentes patogênicos como pequenos genomas[8], como vírus e pequenas bactérias, tem grande parte do seu proteoma[9] sintetizado em um conjunto de peptídeos que podem ser testados experimentalmente. No entanto, para os proteomas maiores, como de parasitas, este tipo de abordagem se torna muito cara e lenta, tornando a identificação de epítopos impraticável, principalmente quando aplicado em grande escala (CARMONA *et al*., 2012).

Neste contexto, a predição in sílico de epítopos de células B tem ganhado bastante espaço e vem se tornado uma alternativa muito interessante, pois permite identificar potenciais candidatos a epítopos antes de aplicar os testes experimentais. Esse fato, contribui de forma significativa para a redução de custos operacionais e de longos períodos de trabalho laboratorial.

2.5 – Predição In Sílico de Epítopos

Atualmente, há diversas abordagens para a predição de epítopos lineares e conformacionais de células B. Para os lineares é preciso apenas uma sequência de aminoácidos da estrutura primária da proteína, enquanto para os conformacionais, é necessário cálculos sobre os dados da estrutura 3D (YANG, YU, 2009). Segundo Davydov e Tonevitsky (2009), a predição de epítopos conformacionais é muito mais complexa, fazendo com que a maioria das abordagens in sílico propostas sejam destinadas para a predição de epítopos lineares.

Os métodos de predição de epítopos lineares de células B podem ser baseados em escala de propensão, baseados em aprendizado de máquina ou híbridos.

O primeiro método in sílico baseado em escala de propensão de aminoácidos foi proposto a exatos 40 anos (HOPP; WOODS, 1981). Após ele vários outros métodos foram propostos e adotavam a mesma estratégia, observavam as propriedades físico-químicas dos aminoácidos que compunham os epítopos, atribuindo-lhes valores para tais propriedades e, em seguida, as analisavam, por meio de uma janela deslizante de tamanhos diferentes (SUN *et al*., 2019). Entretanto, Blyther e Flower (2005), avaliaram os métodos de predição de epítopos lineares de células B, baseados exclusivamente na observação da escala dessas propriedades, combinando cerca de 484 escalas de propensão de aminoácidos, constataram que a combinação das melhores escalas atingia apenas uma predição muito próxima de uma aleatória.

[8] Segundo (GILBERT, 2001), genoma é toda a informação hereditária de um organismo que está codificada em seu DNA (ou, no caso de vírus, no RNA)

[9] Segundo (GARCIA, 1998), proteoma é o conjunto de proteínas e variantes de proteínas que podem ser encontrados em uma célula específica quando sujeita a certos estímulos.

Segundo Greenbaum *et al*. (2007), para melhorar o desempenho da predição de epítopos lineares de células B, deve-se combinar, de forma eficiente, as escalas de propensão com técnicas de aprendizado de máquina.

O princípio básico dos métodos de predição de epítopos lineares de célula B, baseados em aprendizado de máquina, está em coletar um conjunto de dados amplo e variado, pré-processá-los, extrair as características das proteínas dos antígenos, como, por exemplo, propriedades físico-químicas, informações evolutivas e a composição dos aminoácidos, e, ao final, treinar o algoritmo para gerar os modelos de conhecimento.

Este contexto atrelado ao avanço tecnológico da última década e ao número crescente de epítopos lineares de células B caracterizados experimentalmente e disponibilizados em repositórios públicos, estimularam o interesse de muitos autores em começar a utilizar diferentes abordagens de aprendizado de máquina para predição.

3 – TRABALHOS CORRELATOS

O ABCPred (2006) utilizou Redes Neurais Recorrentes (RNN), com 16 neurônio na camada de entrada e 35 neurônios em uma única camada escondida, para prever epítopos lineares de células B. O conjunto de dados foi obtido de Bcipep (SAHA *et al*., 2005) e continha 2479 epítopos positivos de tamanho variável. A fim de padronizar o tamanho dos epítopos e facilitar na aprendizagem da RNN, todos os epítopos redundantes e formados por mais de 20 aminoácidos, foram removidos, resultando em um conjunto final de 700 epítopos positivos. Os epítopos negativos, de comprimento fixado em 20, foram extraídos aleatoriamente de *Swiss-Prot* (BAIROCH; APWEILER, 2000). Foram utilizadas janelas deslizantes variando de 10 a 20 aminoácidos, em volta de cada resíduo, para obter as suas propriedades e, assim, alimentar a RNN. Para os testes, foi utilizada a validação cruzada quíntupla (5-fold *cross validation*), obtendo uma precisão de 66% e MCC de 0,3128, com uma janela deslizante de tamanho 16 (SAHA; RAGHAVA, 2006).

BCPred (2008) utilizou Máquina de Vetor de Suporte com cinco variações de Kernel e validação cruzada quíntupla, para prever epítopos lineares de células B. Utilizou um conjunto de dados reduzido contendo 701 amostras de epítopos positivos, extraídos de Bcipep (SAHA *et al*., 2005) e 701 epítopos negativos, extraídos aleatoriamente de *Swiss-Prot* (BAIROCH; APWEILER, 2000). Ao final dos testes, utilizando o kernel de base radial para o SVM e uma janela deslizante de tamanho 20, para obter as propriedades dos aminoácidos, o método proposto gerou os resultados mais promissores, alcançando uma AUC de aproximadamente 0.76 (EL-MANZALAWY *et al*.,2008).

BayesB (2010) é um método proposto para predizer epítopos lineares de células B baseado na estrutura da proteína. O método utiliza SVM e extração de características Bayes, para prever epítopos de diversos tamanhos (12 a 20). Foram utilizados 2 conjuntos de dados de referência (EL-MANZALAWY *et al*., 2008) e (CHEN *et al*., 2007). A primeira base continha 701 epítopos de comprimento cinco de peptídeos diferentes (12, 14, 16, 18 e 20). A segunda, continha 872 epítopos de tamanho único (20). Em ambos os conjuntos, quantidade iguais de epítopos negativos foram retirados, de forma aleatória, pelos autores, do banco de dados Uniprot (BAIROCH *et al*., 2005). Os vetores de características foram codificados de uma maneira bi-perfil (SHAO *et al*., 2009), contendo atributos de perfil específico de posição positiva e perfil específico de posição negativa. Estes perfis foram gerados calculando a frequência de ocorrência de cada aminoácido em cada posição da sequência do peptídeo no conjunto de epítopos positivos e negativos. Desta forma, cada peptídeo de entrada (tamanho 20,

por exemplo), seria codificado por um vetor de dimensão 40 (20 x 2), na qual contém informações sobre os resíduos nos espaços positivos (epítopos) e nos espaços negativos (não epítopo). Para os testes, foi utilizada a validação cruzada de 10 vezes, obtendo uma precisão de 74.5% (WEE *et al*., 2010).

SVMTrip (2012) é um método de predição de epítopos antigênicos lineares de células B, que combina SVM com a similaridade Tri-peptídeo e pontuações de propensão. O conjunto de dados foi extraído do IEDB (VITA *et al*., 2010) e continha 65456 epítopos positivos. Após agrupá-los de acordo com o grau de similaridade (maior que 30%), medidos pelo BLAST, eliminando os epítopos redundantes, o conjunto de dados resultante consistia de 4925 epítopos positivos. Os negativos foram extraídos de segmentos não epítopos das sequências de antígenos correspondentes, mantendo o mesmo número de subsequências de comprimento. O SVMTriP atingiu uma sensibilidade de 80.1%, uma precisão de 55.2% e o valor de AUC de 0.702, utilizando validação cruzada quíntupla (YAO *et al*., 2012).

BeePro (2013) é um método proposto para predizer epítopos lineares e conformacionais de células B, por meio de informações evolutivas e escalas de propensão. Este método utiliza 16 propriedades para a construção do vetor de característica, nas quais incluem: a Matriz de Pontuação Específica de Posição (PSSM), uma escala de razão de aminoácidos e um conjunto de 14 propriedades físico-químicas, obtidas por meio de um processo de seleção de características. Para o treinamento foi utilizado o SVM com o kernel base radial, validação cruzada quíntupla e 7 conjuntos de dados: (SOLLNER *et al*., 2008), AntiJen1 e AntiJen2 (TOSELAND *et al*., 2005), HIV (KORBER *et al*., 2003; PELLEQUER *et al*., 1993), PC (WANG *et al*., 2011) e Benchmark (PONOMARENKO; BOURNE, 2007), para evitar viés e distorções nos resultados. O BeePRo atingiu uma AUC e uma precisão que variam de 0.9874 a 0.9950 e 93,73% a 97,31%, respectivamente (LIN *et al*., 2013).

SVM BeeproPipe (2015) é uma ferramenta de predição de epítopos lineares de células B, que utiliza diferentes táxons para o treinamento e validação (bactérias, vírus e protozoários). O conjunto de dados utilizado foi extraído de IEDB (VITA *et al*., 2010), sendo utilizado epítopos positivos, negativos e não epítopos. Positivos e negativos são epítopos e sequências que se mostraram não epítopos, respectivamente, após validação experimental. Os não epítopos são sequências de aminoácidos que não foram testadas experimentalmente, mas são derivadas das mesmas proteínas que apresentam os epítopos positivos. Para formar os conjuntos de dados finais, epítopos redundantes foram eliminados por meio da ferramenta BLAST (80% de similaridade). Ao final, totalizou-se 5548 epítopos positivos e 5882 epítopos negativos, criando duas bases de dados, uma com epítopos positivos e negativos e outra com epítopos positivos e

não epítopos, seguindo uma distribuição balanceada em cada classe (50%). Como estratégia de predição, esta ferramenta utilizou 14 propriedades físico-químicas, a taxa de proporção de aminoácidos em cada proteína e o SVM com Kernel radial. Para criar o vetor de característica (embeddings), foi utilizada uma janela deslizante de tamanho 20, centrada em cada aminoácido. Para os testes foi utilizada validação cruzada quíntupla, obtendo uma precisão de 95.71%, 94.90%, 92.01%, 94.94%, 89.09% e 88.74% para os conjuntos de dados de protozoário, bactéria e vírus, respectivamente, e para epítopos positivos/negativos e epítopos positivos/não epítopos, respectivamente (LOPES, 2015).

BepiPred-2.0 (2017) é um Sistema Web utilizado na predição de epítopos de células B, a partir de sequências de antígenos. A base de dados utilizada para treinamento foi obtida do IEDB (VITA *et al*., 2010) e continha 11834 epítopos positivos e 18722 epítopos negativos. Os peptídeos menores do que 5 ou maiores do que 25 aminoácidos foram removidos, pois raramente os epítopos estão fora desse intervalo. O método utilizado para treinamento foi o algoritmo "*Random Forest Regression*" (RF) com validação cruzada quíntupla. Cada resíduo foi codificado utilizando o seu volume, hidrofobicidade, polaridade, acessibilidade de superfície relativa e estrutura secundária, além do volume total do antígeno, gerado pela soma individual de cada resíduo. Na etapa de pré-processamento foi utilizado uma janela deslizante de tamanho 9, centrada em cada resíduo, para percorrer todo o peptídeo. Segundo os seus testes, os autores afirmaram que BepiPred-2.0 superou (na média) todas as principais ferramentas de predição de epítopos lineares de células B, até então disponíveis na literatura (JESPERSEN *et al*., 2017).

EpiDope (2020) é uma ferramenta desenvolvida em Python que utiliza Redes Neurais Profundas (DNN) para detectar regiões de epítopos de células B em sequências individuais de proteínas. O conjunto de dados para treinamento foi extraído do IEDB (VITA *et al*., 2010) e continha, originalmente, 30556 sequências de proteínas, em que cada uma continha epítopos verificados experimentalmente ou epítopos negativos. Para melhor representar o conjunto de dados, foi fundido sequências de proteínas idênticas, porém mantendo as informações sobre suas regiões verificadas (representa um epítopo ou um não epítopo verificado experimentalmente), reduzindo de 30556 para 3186 sequências de proteínas. Em seguida, para eliminar sequências de proteínas semelhantes e reduzir redundância, foram utilizados dois limiares de identidade: 0,8 e 0,5. A arquitetura do EpiDope consiste de duas partes: a primeira, utiliza embeddings sensíveis ao contexto dos aminoácidos, sendo cada um, codificado por um vetor de comprimento 1024, que codifica informações físico-químicas e estruturais. Esses embeddings são previamente calculados por ELMo (HOCHREITER; SCHMIDHUBER, 1997)

e são a entrada para uma camada LSTM bidirecional (2 x 5 nós), seguida por uma camada densa com 10 nós; a segunda, codifica cada aminoácido em um vetor de comprimento 10, não sensível ao contexto, que é conectada a uma camada LSTM (HOCHREITER; SCHMIDHUBER, 1997) bidirecional (2 x 10 nós), seguida de uma camada densa com 10 nós. Ao final da estrutura, ambas as camadas são conectadas a uma camada densa adicional contendo 10 nós, que se liga a uma camada de saída com 2 nós. Para a validação foi utilizado validação cruzada de 10 vezes, obtendo um AUC de 0.605, contra os 0.465, obtidos pela ferramenta BepiPred-2.0, que segundo o autor, era a líder atual na literatura (COLLATZ *et al*., 2020).

Diante dos trabalhos acima mencionados, pode-se destacar alguns problemas intrínsecos associados as técnicas de aprendizado de máquina utilizados, como treinamento *off-line*, demanda de várias épocas para convergência e necessidade de gerar novos modelos de conhecimento à medida que dados de entrada inéditos surgem na literatura (aprimoramento do conhecimento).

Neste contexto, dentre as diversas técnicas de aprendizado de máquina presentes na inteligência artificial, como, por exemplo, SVM, Regressão de Floresta Aleatória, Algoritmos Genéticos, *Naive Bayes* e RNAs, pode-se destacar as RNAs da família ART (*Adptative Ressonance Theory)*. Apesar de ser uma RNA proposta na década de 80 (CARPENTER; GROSSBERG, 1987a), ela é bastante competitiva em relação as demais técnicas, inclusive permite treinamento on-line, converge em poucas épocas e tem a capacidade de aprender novos padrões sem esquecer o que já foi aprendido.

A tabela 1 apresenta, simplificadamente, as principais características das ferramentas/métodos mencionados ao longo deste capítulo.

Tabela 1 - As Principais Características dos Trabalhos Correlatos

Ferramenta	Técnica de Aprendizagem de Máquina Utilizado	Tipo de Abordagem	Conjunto de Dados	Quantidade de Epítopos Positivos/Negativos	Comprimento dos Epítopos Positivos/Negativos	Tamanho da Janela	Tipo de Particionamento de Dados	Métricas Utilizadas
ABCPred	Redes Neurais Recorrentes	Físico-Químicas	BCIPEP	700/700	20	16	Validação Cruzada de 5 vezes	Precisão: 66% MCC: 0,3128
BCPred	SVM	Físico-Químicas	BCIPEP	701/701	12, 14, 16 e 18	20	Validação Cruzada de 5 vezes	AUC ROC: 76%
BayesB	SVM e extração de características Bayes	Atributos de perfil específico de posição positiva e perfil específico de posição negativa	BCIPEP/ CHEN872	701/701 872/872	12, 14, 16, 18 e 20	20	Validação Cruzada de 10 vezes	Precisão: 74,5%
SVMTrip	SVM	Similaridade Tri-peptídeo e Pontuações de Propensão	IEDB	4925/4925	10, 12, 14, 16, 18 e 20	10, 12, 14, 16, 18 e 20	Validação Cruzada de 5 vezes	Sensibilidade: 80.1% Precisão: 55,2% AUC ROC: 0,702
BeePro	SVM	Físico-Química PSSM Proporção de aminoácidos	Sollner AntiJen1 AntiJen2 HIV Pellequer PC Benchmark	112/111 309/309 691/689 65/65 78/77 98/98 -	-	20	Validação Cruzada de 5 vezes	Variam entre os conjuntos de dados. AUC ROC: 0,9874 a 0,9950 Precisão: 93,73% a 97,31%
SVM BeeproPipe	SVM	Físico-Química Proporção de aminoácidos	IEDB	5548/5882	-	20	Validação Cruzada de 5 vezes	Aplicado por táxon (Protozoário, Bactéria e Vírus) Precisão: 95,71%, 94,90% e 92,01%

BepiPred-2.0	*Randon Forest Regression*	Físico-Químicas	IEDB	11834/18722	5 – 25	9	Validação Cruzada de 5 vezes	AUC ROC: 0,62
EpiDope	Rede Neural Recorrente (LSTM) mais Rede Neural Artificial Tradicional (Densa)	Físico-Químicas Informações Estruturais	IEDB	8519/16091	13	-	Validação Cruzada de 10 vezes	AUC ROC: 0,67 AUC Precision-Recall:

Fonte: Próprio Autor

4 – TEORIA DA RESSONÂNCIA ADAPTATIVA

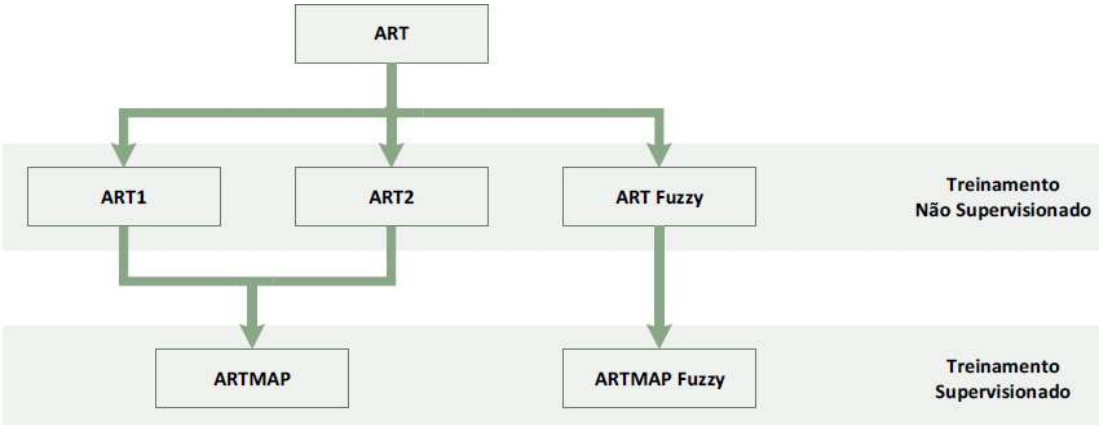
Com os avanços tecnológicos dos últimos anos, uma das áreas que tem se mostrado bastante promissora para a solução de problemas em diversos campos é a Inteligência Artificial. Um ramo é formado pelas Redes neurais artificiais, que ganharam bastante espaço em vários tipos de aplicações, tais como: diagnósticos, processamento de imagens, previsão, classificação, reconhecimento de padrões, otimização, dentre outros. A fim de atender aos mais variados tipos de problemas e situações, diferentes tipos de arquiteturas de redes neurais foram propostos. Uma que se destacou consideravelmente, dentre as arquiteturas existentes, foram as redes neurais baseadas na teoria da ressonância adaptativa (do inglês *Adptative Ressonance Theory* – ART) (CARPENTER; GROSSBERG, 1987a).

Carpenter e Grossberg (1987a) foram os pioneiros das arquiteturas iniciais das redes neurais da família ART. Elas permitiam uma aprendizagem estável, rápida e incremental em ambientes que variavam no tempo, generalização múltipla e convergência rápida com um número relativamente pequeno de padrões de treinamento. Estas características foram realçadas, devido aos autores buscarem pelo desenvolvimento de um modelo de rede, capaz de prezar pela plasticidade e pela estabilidade, ou seja, grande facilidade de aprender continuamente novos padrões, sem perder a memória dos padrões já aprendidos.

A plasticidade e a estabilidade contribuíram diretamente para o desenvolvimento de uma estrutura neural capaz de ser adaptável em repostas às novas informações, ainda que relevantes, e capaz de preservar os conhecimentos já adquiridos, além de flexível para o armazenamento de novas informações (GROSSBERG, 2013; LOPES *et al*., 2005).

A Figura 5 apresenta os principais modelos que fazem parte da família ART.

Figura 5 - Família ART



Fonte: (SANTOS JÚNIOR 2017)

As redes neurais ART_1, ART_2 e ART FUZZY, possuem treinamento não supervisionado e a capacidade de reconhecer padrões de entrada de forma arbitrária. No entanto, a rede ART_1 reconhece apenas padrões binários, enquanto que a rede ART_2, reconhece tanto binários quanto discretos/contínuos (CARPENTER; GROSSBERG, 1987a) (CARPENTER; GROSSBERG, 1987b). Já o diferencial acrescentado na arquitetura da rede ART FUZZY é o fato de utilizar cálculos baseados na lógica fuzzy (CARPENTER *et al*., 1991a).

As redes neurais ARTMAP e ARTMAP-FUZZY, possuem treinamento supervisionado e são compostas por dois módulos ART interconectados, por meio do módulo Inter-ART. Ambas identificam padrões de entradas binários e discretos/contínuos, no entanto, a ARTMAP-FUZZY, acrescenta a todos os seus cálculos os fundamentos da lógica fuzzy (CARPENTER *et al*., 1991b).

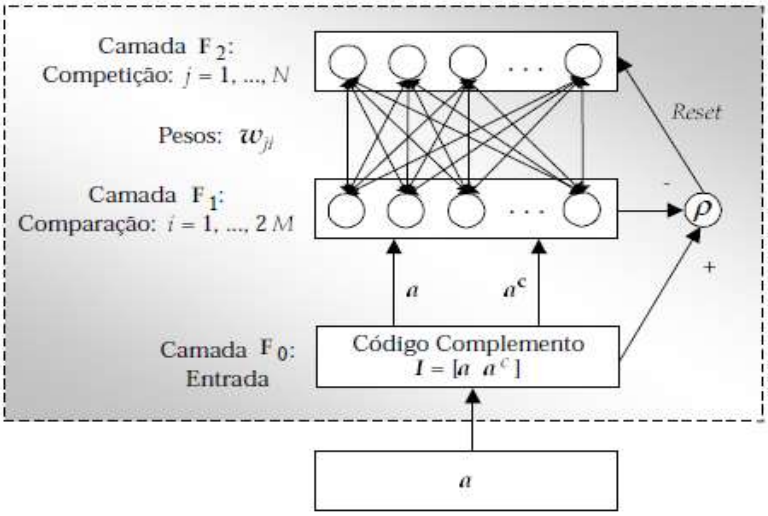
4.1 – Rede Neural ART-FUZZY

A rede ART FUZZY é baseada em um sistema de aprendizado competitivo que auto-organiza as categorias de maneira estável, à medida que sequências de padrões de entrada analógicos, compreendidos entre 0 e 1 se formam. O princípio desses grupos é formar regras abstratas sobre a distribuição dos dados, para garantir a generalização do modelo (CARPENTER *et al*., 1992).

A rapidez da aprendizagem desse algoritmo não supervisionado está atrelada ao uso dos operadores fuzzy MIN (∧) e MAX (∨), advindos da teoria da lógica fuzzy, combinado com a codificação do complemento, que garante a amplitude da informação e com a normalização dos vetores de entrada (LOPES, 2005; CARPENTER *et al*., 1992).

A arquitetura da rede ART-FUZZY é composta por três camadas: F_0, F_1 e F_2, responsáveis por verificarem a semelhança de um novo padrão apresentado à rede com um já representado por um neurônio. Há sinais de controle em cada camada, responsáveis por garantirem as interconexões entre as camadas e a manipulação do fluxo dos dados. As camadas F_1 e F_2 são interconectadas por meio de pesos *feedfoward* (w_i) e *feedbackward* (w_j), que garantem o armazenamento das informações (CARPENTER *et al*., 1992). A Figura 6 apresenta a arquitetura geral da rede ART-FUZZY.

Figura 6 - Arquitetura da Rede Neural ART FUZZY



Fonte: (LOPES, 2005)

4.1.1 – A Camada F_0

A camada de entrada F_0 é responsável por normalizar e codificar em complemento o vetor de entrada *a,* um número real que pertence ao intervalo [0, 1] de dimensão *M*, para gerar o vetor de atividade **I**. Em (1) é apresentado a operação de normalização e em (2) a função norma (CARPENTER *et al*., 1992).

$$\boldsymbol{I} = \frac{\boldsymbol{a}}{|\boldsymbol{a}|} \tag{1}$$

$$|\boldsymbol{p}| = \sum_{i=1}^{M} |p_i| \tag{2}$$

sendo:

$\boldsymbol{a}$ = vetor de entrada

$\boldsymbol{I} = [I_1, I_2, I_3, \ldots, I_M]$

| . | = função norma

O processo de codificação em complemento gera o vetor de entrada $\boldsymbol{I}$ com *2M* elementos, conforme definido em (3) (CARPENTER *et al*., 1992).

$$\boldsymbol{I} = (\boldsymbol{a}, \boldsymbol{a^c}) = (a_1, a_2, \ldots, a_M, a_1^c, a_2^c, \ldots, a_M^c) \tag{3}$$

sendo:

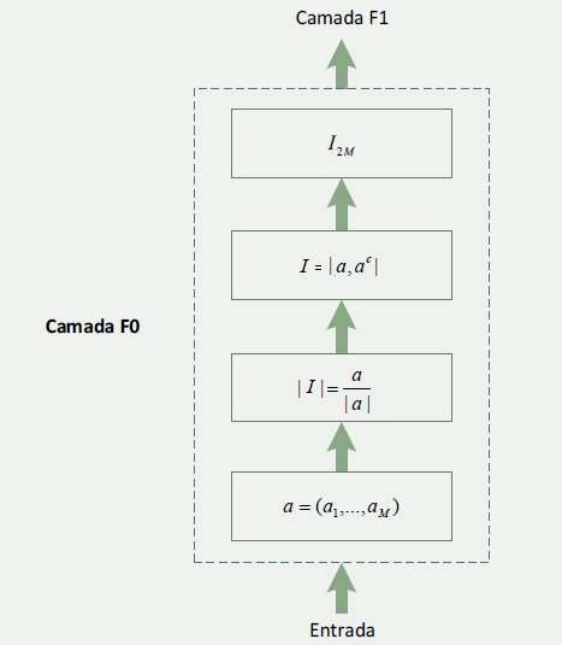
$$a_1^c = 1 - a_i$$

Em seguida o vetor de atividade $\boldsymbol{I}$ gerado é enviado à camada superior F_1. Vale ressaltar, que todo esse processo preserva a amplitude do padrão de entrada, conforme demonstrado em (4) (CARPENTER *et al.*, 1992).

$$\boldsymbol{I} = |\boldsymbol{a}, \boldsymbol{a^c}| = \sum_{i=1}^{M} a_i + (M - \sum_{i=1}^{M} a_i) = M \tag{4}$$

A Figura 7 apresenta a sequência de etapas da camada F_0, conforme descritos anteriormente.

Figura 7 - Etapas da camada F_0



Fonte: SANTOS JÚNIOR, 2017

4.1.2 – A Camada F_1

A camada de comparação F_1 recebe sinais da camada F_0 (o vetor de atividade $\boldsymbol{I}$) e sinais da camada F_2 (o vetor de ativação), denotado pelo vetor $\mathbf{y} = [\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_\mathrm{N}]$ (CARPENTER *et al.*, 1992).

A operação fuzzy AND, denotada pela equação (6) (CARPENTER *et al.*, 1992), aplicada entre o vetor de atividade $\boldsymbol{I}$ e o vetor de pesos $\boldsymbol{w_j}$, referente à categoria ativa J da

camada F_2, resulta no vetor de ativação $\boldsymbol{x} = [x_1, x_2, \ldots, x_{2M}]$ da camada F_1, na qual atende a equação (7) (CARPENTER *et al*., 1992).

$$(p \wedge q)_i = min(p_i, q_i) \tag{6}$$

$$x = \begin{cases} \boldsymbol{I}, & se\ F_2\ estiver\ inativa \\ \boldsymbol{I} \wedge \boldsymbol{w_j}, & se\ existir\ categoria\ ativa\ em\ F_2 \end{cases} \tag{7}$$

sendo:

***I*:** vetor de ativação da camada F_0;

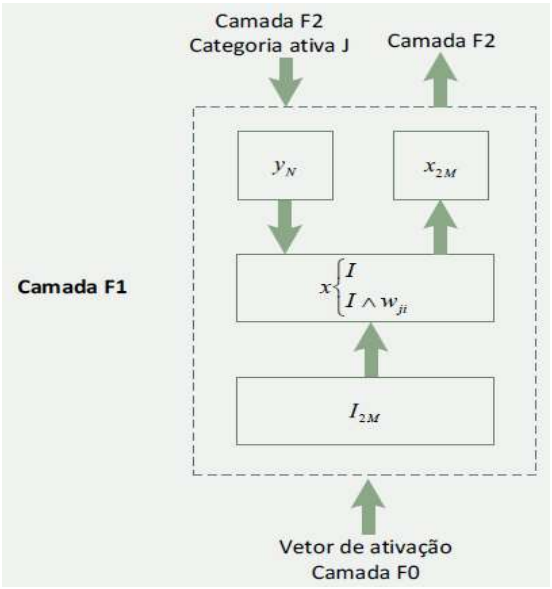
$\boldsymbol{w_j}$: vetor de pesos entre as camadas F_1 *e* F_2;

^: operador fuzzy AND.

A memória da rede ART FUZZY é representada pela matriz peso $\boldsymbol{w_{ij}}$, que se localiza entre as camadas F_1 *e* F_2. O índice *i* de dimensão *2M*, representa cada elemento da categoria armazenada e o índice *j* de dimensão *N*, representa cada categoria. É importante mencionar que a dimensão *N* aumenta à medida que uma nova categoria é armazenada na matriz de pesos ***w*** (CARPENTER *et al*., 1992).

A Figura 8 ilustra a sequência das etapas supracitadas da camada F_1.

Figura 8 - Etapas da Camada F_1



Fonte: (SANTOS JÚNIOR, 2017)

4.1.3 – A Camada F_2

A camada de reconhecimento F_2 é responsável por classificar os padrões de treinamento em categorias de reconhecimento aprendidas pela rede, por meio de seus *N* neurônios. Para isso, a camada F_2 seleciona e ativa uma categoria *J* e a envia à camada F_1, por meio de seu vetor de ativação $\boldsymbol{y} = [y_1, y_2, \ldots, y_N]$ e de seu vetor de pesos $\boldsymbol{w_j}$, que corresponde ao neurônio vencedor selecionado *(J)* (CARPENTER *et al*., 1992).

Para o processo de escolha da categoria *J* é utilizado a função T_j, definida pela equação (8) (CARPENTER *et al*., 1992), que atribui um valor à cada neurônio da camada F_2, e somente aquele que possuir o maior valor é selecionado. Caso exista mais de uma categoria *J* ativa, é selecionado o neurônio *J* de menor índice (CARPENTER *et al*., 1992).

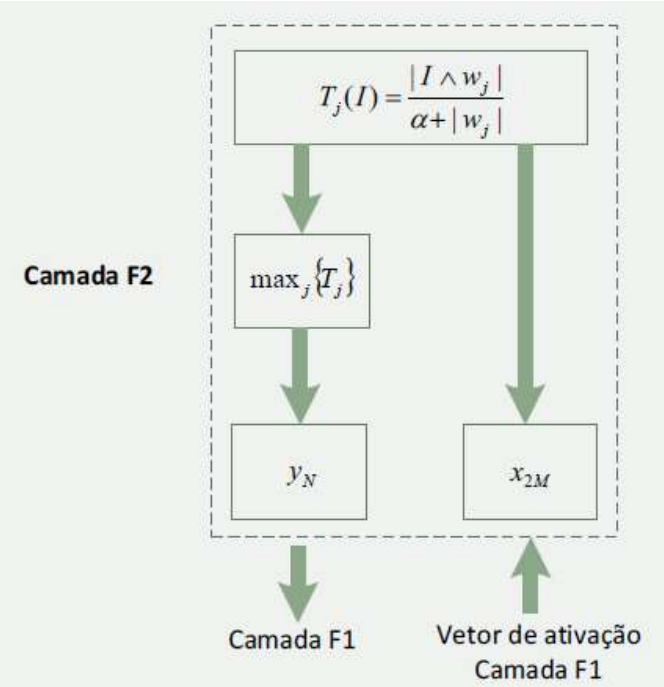
$$T_j(I) = \frac{|\boldsymbol{I} \wedge \boldsymbol{w_j}|}{\alpha + |\boldsymbol{w_j}|} \tag{8}$$

sendo:

α: parâmetro de escolha ($\alpha \geq 0$).

A Figura 9 ilustra o processo da camada de reconhecimento F_2.

Figura 9 - Etapas da camada F_2



Fonte: SANTOS JÚNIOR, 2017

4.1.4 – Ressonância

O estado de ressonância ou reset são ativados de acordo com o grau de semelhança entre o padrão de entrada e a categoria *J* ativa. Caso a semelhança entre eles atinja a exigida pela rede neural, ocorre a ressonância, caso contrário, o reset. Os neurônios que não passam no teste de semelhança são desabilitados até o final do processo atual, e isso se repete, até que se encontre um neurônio que seja aprovado. Caso nenhum neurônio seja qualificado, uma nova categoria na camada F_2 é criada e associada ao padrão de entrada corrente (CARPENTER *et al*., 1992).

O parâmetro de vigilância ρ define o grau mínimo de semelhança que deve ser atingido no teste para que ocorra a ressonância, sendo $\rho \in [0, 1]$. Caso isso aconteça, o padrão de entrada é incluído na categoria ativa para consequente processo de aprendizagem, caso contrário, o reset é aplicado. O teste de semelhança é definido pela equação (9) (CARPENTER *et al*., 1992).

$$\frac{|\boldsymbol{I} \wedge \boldsymbol{w_j}|}{|\boldsymbol{I}|} \geq \rho \tag{9}$$

sendo:

$\boldsymbol{I}$: vetor de ativação da camada F_0;

$\boldsymbol{w_j}$: vetor de pesos da categoria ativa;

ρ: parâmetro de vigilância.

4.1.5 – Processo de Aprendizagem

Concluído o processo de escolha de categoria para todos os vetores de ativação, o processo de atualização dos pesos da matriz $\boldsymbol{w_j}$ se inicia, de acordo com a equação (10) (CARPENTER *et al*., 1992).

$$\boldsymbol{w_J^{novo}} = \beta * (\boldsymbol{I} \wedge \boldsymbol{w_J^{velho}}) + (1 - \beta) * \boldsymbol{w_J^{velho}} \tag{10}$$

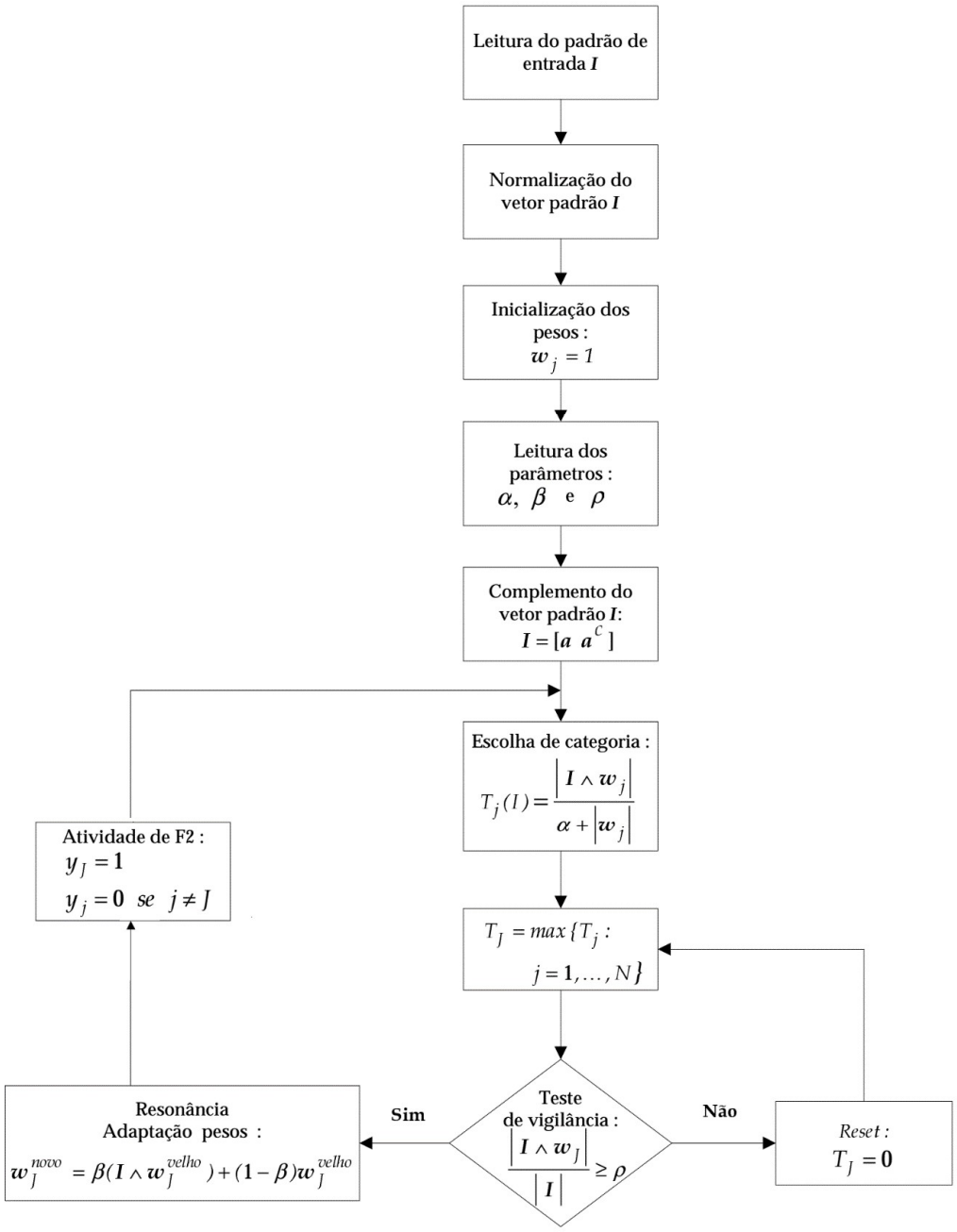
sendo:

β: é a velocidade de aprendizagem (taxa de treinamento) e pertence a $[0, 1]$.

Para o processo de aprendizagem há dois possíveis tipos de treinamento: o treinamento lento e o treinamento rápido. No treinamento rápido, quando $\beta = 1$, geralmente os pesos são ajustados para seus valores ótimos em apenas um ciclo de treinamento (época). Já no treinamento lento, quando $\beta < 1$, os pesos são ajustados lentamente em várias épocas (CARPENTER *et al*., 1991a; LOPES, 2005).

A Figura 10 ilustra o fluxograma que apresenta todo esse processo de aprendizagem da RNA ART FUZZY.

Figura 10 - Fluxograma da Rede Neural ART FUZZY

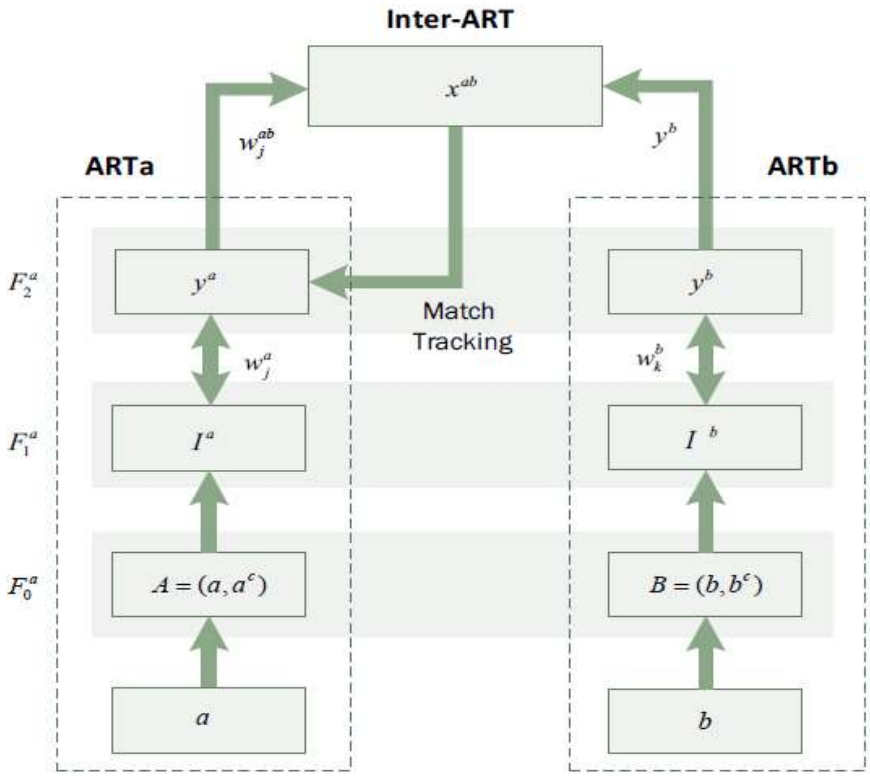


Fonte: LOPES, 2005

4.2 – Rede Neural ARTMAP-FUZZY

A RNA ARTMAP-FUZZY utiliza aprendizagem supervisionada e conceitos da teoria dos conjuntos em seus cálculos, sendo capaz de aprender categorias estáveis em resposta a padrões de entradas binários ou discreto/contínuos. O modelo é composto por dois módulos ART FUZZY, denominados ART_a e ART_b e um módulo Inter-ART (CARPENTER *et al.*, 1992). A Figura 11 apresenta a arquitetura da rede ARTMAP-FUZZY.

Figura 11 - Arquitetura ARTMAP-FUZZY



Fonte: (SANTOS JÚNIOR, 2017)

Conforme ilustra a Figura 11, os padrões de entrada são apresentados ao módulo ART_a, enquanto que os padrões de saída são apresentados ao módulo ART_b. Já o módulo Inter-ART, estabelece o mapeamento entre ambos os padrões (CARPENTER *et al.*, 1992).

4.2.1 – Módulo INTER-ART

O módulo Inter-ART verifica a correspondência entre as categorias ativas dos módulos ART_a e ART_b, criando um vetor de ativação denominado $\boldsymbol{x^{ab}}$, denotado pela equação (11) (CARPENTER *et al.*, 1992).

$$\begin{cases} \boldsymbol{y^b} \wedge \boldsymbol{w_j^{ab}} & se\ F_2^a\ ativo\ e\ F_2^b\ ativo \\ \boldsymbol{w_j^{ab}} & se\ F_2^a\ ativo\ e\ F_2^b\ inativo \\ \boldsymbol{y^b} & se\ F_2^a\ inativo\ e\ F_2^b\ ativo \\ 0 & se\ F_2^a\ inativo\ e\ F_2^b\ inativo \end{cases} \tag{11}$$

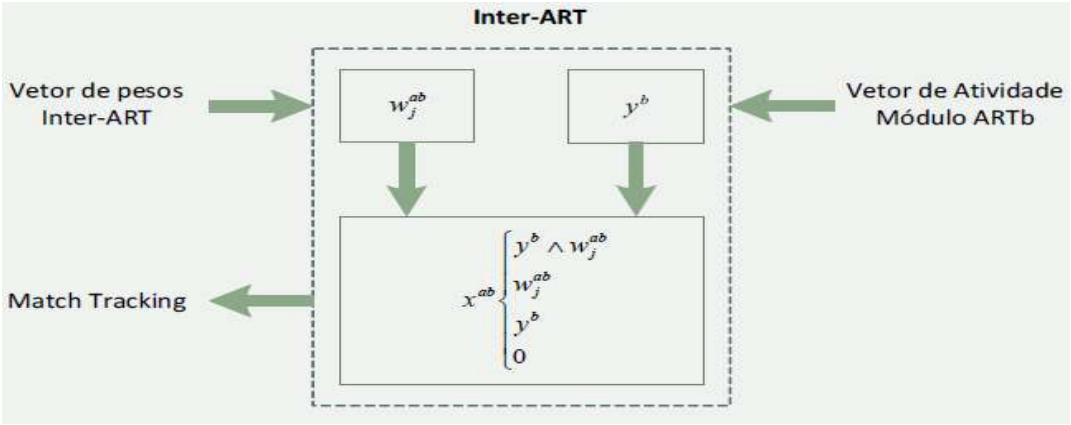
sendo:

$\boldsymbol{y_b}$: vetor de ativação da camada F_2^b do módulo ART_b;

$\boldsymbol{w_j^{ab}}$: vetor do índice J da matriz de pesos do módulo Inter-ART.

O módulo Inter-ART mapeia as categorias do módulo ART_a com sua respectiva categoria em ART_b, via matriz de pesos $\boldsymbol{w^{ab}}$. Após o mapeamento das categorias, o módulo Inter-ART verifica se há correspondência entre elas, por meio do vetor $\boldsymbol{w_j^{ab}}$. Caso não exista, o vetor $\boldsymbol{x^{ab}}$ assume o valor 0 e o processo de busca por categorias no módulo ART_a se repete, até que o critério de equivalência entre o padrão de entrada e o de saída seja satisfeito (CARPENTER *et al*., 1992). A Figura 12 apresenta esse fluxo de atividades do módulo Inter-ART.

Figura 12 - Módulo Inter-ART



Fonte: (SANTOS JÚNIOR, 2017)

4.2.2 – Match Tracking

O *Match Tracking* é um mecanismo interno da rede, autorregulador, que tenta maximizar a generalização e minimizar o erro. À medida que a rede faz um diagnóstico errado, mesmo que a categoria ativa de ART_a satisfaça o teste de vigilância, o parâmetro de vigilância ρ_a é incrementado em uma quantidade mínima estipulada pelo parâmetro ε, a fim de forçar a

rede a escolher outra categoria no módulo ART_a e corrigir o erro no módulo ART_b (CARPENTER *et al*., 1992; LOPES, 2005).

Para o processo de ressonância de ART_a o parâmetro de vigilância ρ_a mantém sempre o seu valor base, denotado $\bar{\rho}_a$. No entanto, no processo de *Match Tracking*, se a equação (12) for satisfeita, ρ_a é ligeiramente incrementado, de acordo com a equação (13), até que satisfaça a equação (14) (CARPENTER *et al*., 1992). Após o processo de *Match Tracking*, o parâmetro ρ_a retorna ao seu valor base.

$$\left|\boldsymbol{x^{ab}}\right| < \rho_{ab} * \left|\boldsymbol{y^{b}}\right| \tag{12}$$

sendo:

ρ_{ab}: parâmetro de vigilância do módulo Inter-ART

$$\rho_a = \frac{\left|\boldsymbol{I^a} \wedge \boldsymbol{w_j^a}\right|}{|\boldsymbol{I^a}|} + \varepsilon \tag{13}$$

sendo:

ε: parâmetro de incremento de ρ_a

$$\left|\boldsymbol{I^a} \wedge \boldsymbol{w_j^a}\right| > \rho_a * |\boldsymbol{I^a}| \tag{14}$$

A busca por uma nova categoria para o padrão de entrada atual se repete até que se encontre um diagnóstico correto ou que se crie uma nova categoria em ART_a e a associe à categoria atual no módulo ART_b (LOPES, 2005).

Após todo esse processo a rede já pode adquirir novas informações, através da atualização dos pesos dos módulos ART_a ($\boldsymbol{w_j^a}$) e Inter-ART ($\boldsymbol{w_{jk}^{ab}}$), conforme denotado pelas equações (15) e (16), respectivamente (CARPENTER *et al*., 1992).

$$\boldsymbol{w_J^{novo}} = \beta * (I \wedge \boldsymbol{w_J^a}) + (1 - \beta) * \boldsymbol{w_J^a} \tag{15}$$

$$\boldsymbol{w_{JK}^{novo}} = 1$$
$$\boldsymbol{w_{jk}^{novo}} = 0 \; para \; j \neq J \; e \; k \neq K \tag{16}$$

sendo:

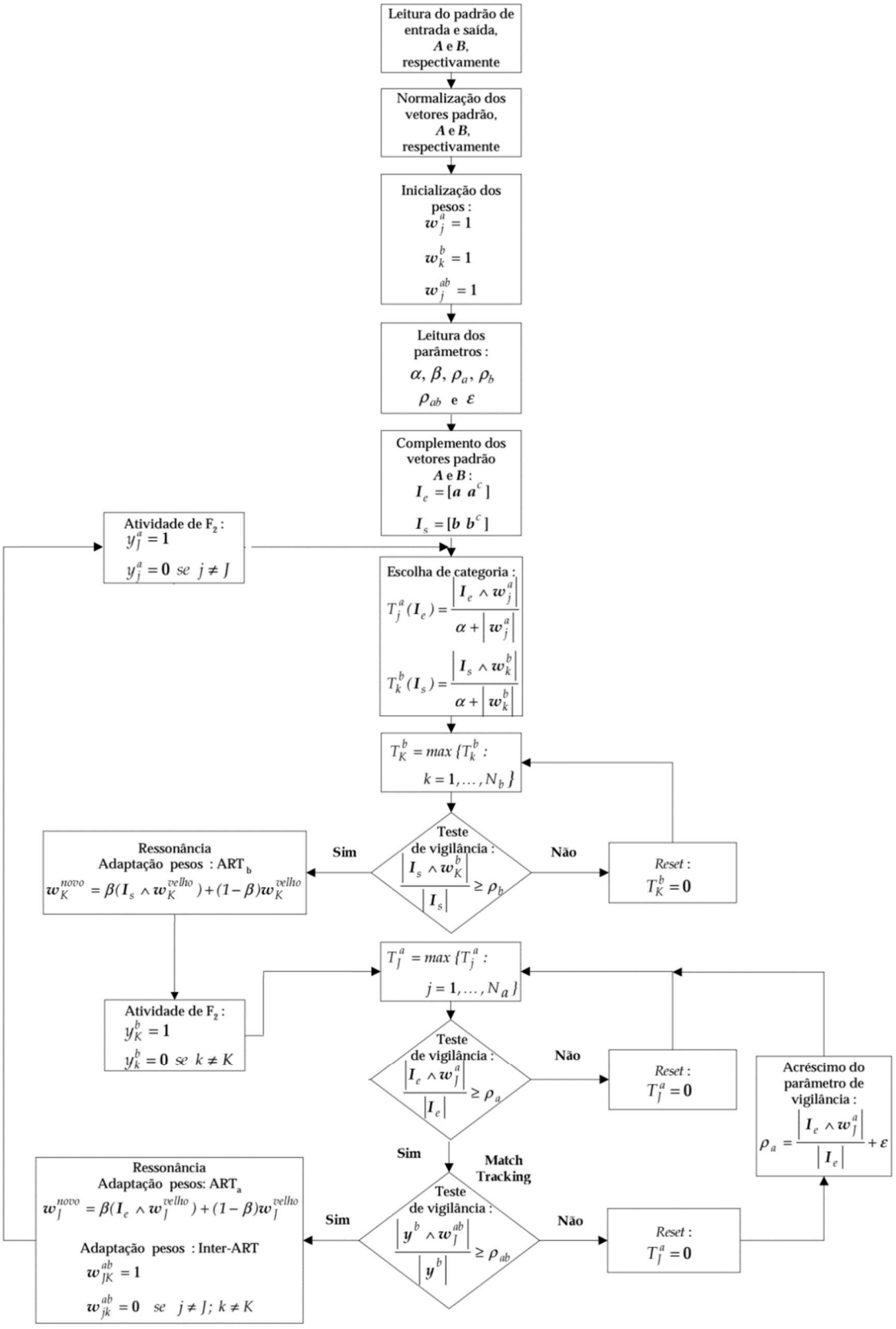
β: taxa de treinamento e $\beta \in [0, 1]$.

Visto que a categoria ativa em ART_b não é mais atualizada até a próxima entrada, a atualização de seus pesos pode ser feita logo após a sua ressonância, conforme a equação (17) (CARPENTER *et al*., 1992).

$$\boldsymbol{w_K^{novo}} = \beta * (I \wedge \boldsymbol{w_K^b}) + (1 - \beta) * \boldsymbol{w_K^b} \tag{17}$$

A Figura 13 ilustra o fluxograma que apresenta todo esse processo de aprendizagem da RNA ARTMAP-FUZZY.

Figura 13 - Fluxograma da Rede Neural ARTMAP-FUZZY

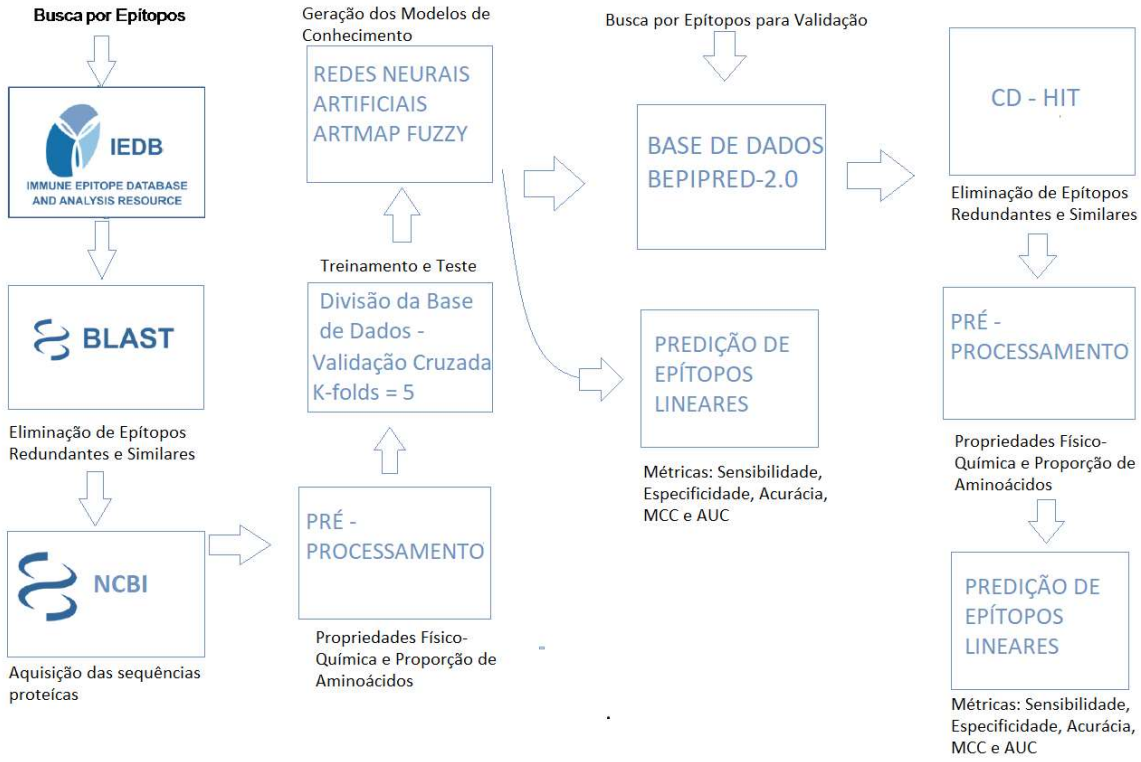


Fonte: (LOPES, 2005)

5 – MATERIAIS E MÉTODOS

Esta seção apresenta os materiais e a metodologia utilizada para a elaboração da ferramenta, além das validações realizadas. A Figura 14 apresenta um esquema simplificado e as seções subsequentes, descrevem cada etapa.

Figura 14 - Esquema da metodologia utilizada no trabalho



Fonte: Próprio Autor

5.1 – Conjunto de Dados de Treinamento e Validação

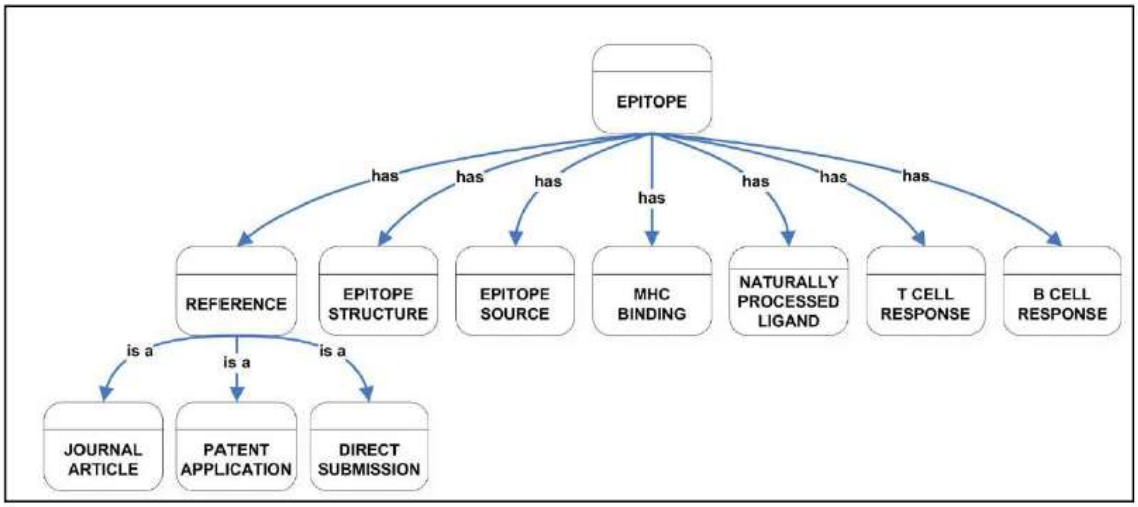
A primeira etapa desta pesquisa consistiu na busca de epítopos lineares de células B, disponíveis em base de dados públicas. Devido a quantidade, a diversidade, a validação experimental e a diversos filtros refinados de pesquisas disponíveis no banco de dados do IEDB, ele foi o escolhido a fornecer tanto os epítopos positivos quanto os negativos aos experimentos de treinamento e teste.

São considerados epítopos positivos os que foram validados experimentalmente e apresentam ensaios (resultados) positivos em pelo menos um experimento. Já os negativos,

também foram validados experimentalmente e não devem apresentar ensaio positivo em nenhum experimento.

A base de dados do IEDB, além de fornecer os epítopos positivos/negativos validados experimentalmente, fornece outras informações importantes (metadados), tais como: o nome e o código do antígeno, a posição de início e de fim do epítopo, o nome e código do organismo, dentre outros. A Figura 15 ilustra a estrutura hierárquica da base de dados do IEDB.

Figura 15 - Visão Geral da Estrutura Hierárquica do IEDB



Fonte: (SATHIAMURTHY *et al*., 2005)

A obtenção dos epítopos positivos e negativos do IEDB foi realizada em 05 de abril de 2021, sendo utilizados os seguintes filtros para a consulta:

- Epítopos lineares;
- Doenças infecciosas de humanos;
- Ensaios de células B;
- Sem restrição de MHC; e
- Bactérias, Vírus e Protozoários.

A Tabela 2 detalha todos os filtros de consulta utilizados no IEDB para a extração dos dados.

Tabela 2 - Filtros de consultas do IEDB

Táxon	Filtro
Bactérias	Nome da doença é infecção bacteriana [DOID:104] Estado/status da doença do hospedeiro selecionado Células T excluídas

	Ligante MHC excluído Organismo alvo é Homo sapiens (humanos)
Protozoários	Nome da doença é Leishmaniose Visceral [DOID: 9146] ou Nome da doença é doença Chagas [DOID: 12140] ou Nome da doença é Cardiomiopatia Chagásica [DOID: 12569] ou Nome da doença é Leishmaniose Cutânea [DOID: 9111] ou Nome da doença é Amebíase [DOID: 9181] ou Nome da doença é Giardíase [DOID: 10718] ou Nome da doença é Leishmaniose [DOID: 9065] ou Nome da doença é Leishmaniose Mucocutânea [DOID: 9155] ou Nome da doença é Malária *Plasmodium Falciparum* [DOID: 14067] ou Nome da doença é Malária *Plasmodium vivax* [DOID: 12978] ou Nome da doença é Taxoplasmose [DOID: 9965] ou Nome da doença é Tricomoníase [DOID: 1947] ou Nome da doença é Tripanossomíase [DOID: 10113] ou Nome da doença é Criptosporidiose [DOID: 1733] ou Estado/status da doença do hospedeiro selecionado Células T excluídas Ligante MHC excluído Organismo alvo é Homo sapiens (humanos)
Vírus	Nome da doença é infecção viral [DOID: 934] Estado/status da doença do hospedeiro selecionado Células T excluídas Ligante MHC excluído Organismo alvo é Homo sapiens (humanos)

Fonte: Próprio Autor

Ao final, a base de dados continha 11509 epítopos positivos e 28080 epítopos negativos, conforme descrito pela Tabela 3.

Tabela 3 - Quantidade de Epítopos Positivos/Negativos extraídos do IEDB

Táxon	Epítopos Positivos	Epítopos Negativos
Bactéria	1803 (15,67%)	4167 (14,84%)
Vírus	5569 (48,39%)	6638 (23,64%)
Protozoário	4137 (35,94%)	17275 (61,52%)
Total	11509	28080

Fonte: Próprio Autor

Foram criadas quatro bases de dados, três independentes, formada cada uma por resíduos de um único táxon, e uma dependente, formada pela união dos resíduos de todos os 3 táxons. A quantidade de epítopos positivos e negativos de cada base é apresentada na Tabela 3 e são nomeadas da seguinte maneira:

- DB_bac: composta apenas por resíduos de epítopos positivos e negativos de bactérias;
- DB_vir: composta apenas por resíduos de epítopos positivos e negativos de vírus;
- DB_prot: composta apenas por resíduos de epítopos positivos e negativos de protozoários; e
- DB_all: composta por resíduos de epítopos positivos e negativos de todos os 3 táxons.

Um dado muito importante que faltava nos metadados da base de dados do IEDB é a sequência da proteína completa de cada antígeno. Pelo fato de os algoritmos de aprendizado de máquina utilizarem o contexto dos aminoácidos da proteína para tentarem encontrar correlações e predizerem possíveis epítopos, fez-se necessário a obtenção desta informação. De posse do código de cada antígeno, obtido da base de dados do IEDB, a base de dados NCBI (NCBI, 2021) foi acessada para recuperar esta informação. Como haviam muitas proteínas para serem recuperadas, foi desenvolvido um *Web Scraping*[10] para realizar esta tarefa.

5.2 – Conjunto de Dados Teste

O conjunto de dados teste foi obtido de (JESPERSEN *et al*., 2017), contendo 30556 sequências de proteína, em que cada sequência possui uma região marcada, que representa um epítopo positivo ou um não epítopo, ambos validados experimentalmente.

Dentre as 30556 sequências, 11834 são epítopos positivos e 18722 são epítopos negativos. O subconjunto de epítopos positivos tem um comprimento médio de 13.99 (quantidade de aminoácidos consecutivos), enquanto que o subconjunto de epítopos negativos tem um comprimento médio de 13.20, conforme apresenta a Tabela 4 (COLLATZ *et al*., 2020).

Tabela 4 - Base de Dados Original

	Base de Dados Original
Regiões Verificadas	30556
Epítopos Positivos	11834
Epítopos Negativos	18722

[10] Segundo (ZHAO, 2017), Web Scraping é uma forma de mineração que permite a extração de dados de sites da web, convertendo-os em informação estruturada para posterior análise.

Comprimento Mediano	15
Média de Comprimento	13,50
Média de Comprimento de Epítopos Positivos	13,99
Média de Comprimento de Epítopos Negativos	13,20

Fonte: Adaptado (COLLATZ *et al*., 2020)

5.3 – Preparação dos Dados de Treinamento e Validação

Os epítopos que apresentavam similaridade maior ou igual a 80% (GAO *et al*., 2012), avaliadas pelo software BLAST (CAMACHO *et al*., 2013), foram agrupados e apenas um de cada grupo foi selecionado aleatoriamente e mantido como uma sequência de epítopo no conjunto de dados final. Esse procedimento é necessário, para evitar que o algoritmo de aprendizado de máquina memorize sequências de epítopos muito similares e favoreça a generalização.

Inicialmente foi necessário realizar alguns ajustes preliminares, com o intuito de eliminar dados irrelevantes e faltantes (valores nulos dos atributos selecionados) e balancear as classes. Foram mantidas apenas as informações dos epítopos, como a sequência de aminoácidos e o código do antígeno, a posição inicial e final do epítopo na proteína e a proteína completa do antígeno.

Para minimizar o desbalanceamento entre as classes e evitar que os modelos de conhecimento gerados sejam tendenciosos e de que epítopos importantes sejam descartados, técnicas de correção de prevalência foram aplicadas ponderadamente, como o de Amostragem Estratificada (GOLDSCHMIDT; PASSOS, 2005), por exemplo. Outros registros removidos foram os que apresentavam epítopos de comprimento superior a 30 e inferior a 5 aminoácidos, pois apareciam esporadicamente no conjunto de dados. Ao final, a base de dados continha 9968 epítopos positivos e 10766 epítopos negativos, conforme descrito, por táxon, pela Tabela 5.

Tabela 5 - Quantidade de Epítopos Positivos/Negativos de cada Táxon após o Pré-Processamento

Táxon	Epítopos Positivos	Epítopos Negativos
Bactéria	1600 (16.05%)	2300 (21.36%)
Vírus	4376 (43.90%)	4475 (41.57%)
Protozoário	3992 (40.05%)	3991 (37.07%)

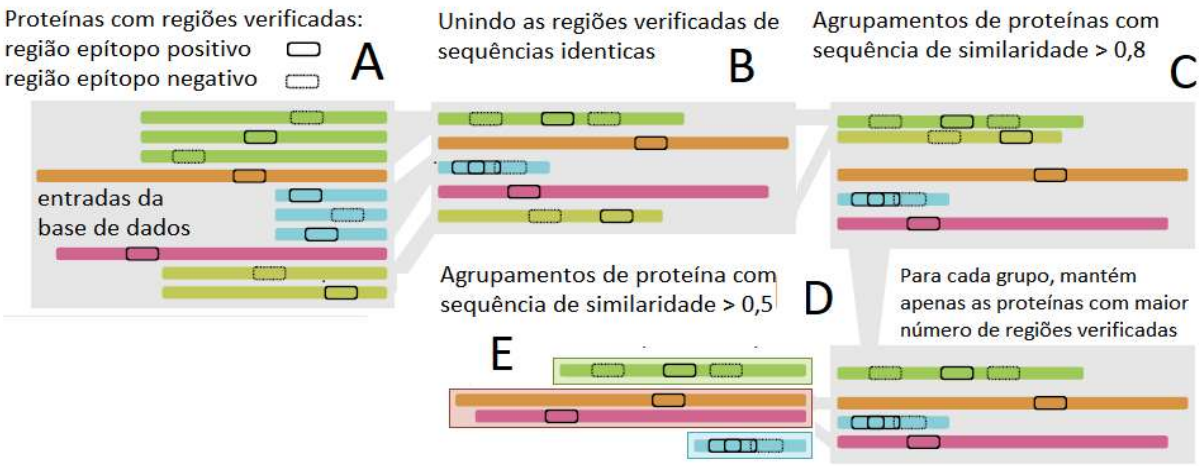
Total	9968	10766

Fonte: Próprio Autor

5.4 – Preparação dos Dados de Teste

A partir do conjunto de dados brutos de (JESPERSEN *et al*., 2017), foram aplicadas todas as etapas de pré-processamento descritas por (COLLATZ *et al*., 2020), a fim de gerar um banco de dados idêntico ao utilizado, e assim, poder comparar os resultados entre os estudos. A Figura 16 apresenta todo esse processo de pré-processamento aplicado no referido conjunto de dados.

Figura 16 - Pré processamento dos dados para gerar o conjunto de dados de teste



Fonte: Adaptado (COLLATZ *et al*., 2020)

Primeiramente, foram unidas as sequências de aminoácidos de proteínas idênticas, preservando as informações sobre suas regiões verificadas (30556) (Figura 16 A e B), resultando em um conjunto de dados reduzido de 3158 proteínas. Em seguida, para atenuar a redundância de sequências de aminoácidos muito similares, foram gerados agrupamentos de todas as sequências com a ferramenta CD-HIT (LI; GODZIK, 2006), utilizando um limiar de 0.8 (Figura 16 C). Foram gerados 1798 grupos de sequências, em que foi recuperada, de cada grupo, apenas a sequência que continha o maior número de regiões verificadas (Figura 16 D), resultando em 24610 regiões. A etapa de agrupamento foi novamente aplicada sobre o conjunto de dados reduzido, porém utilizando um limiar de 0.5 (Figura 16 E), resultando em 1378 grupos de sequências. Novamente, foi recuperado, de cada grupo, apenas a sequência que continha o

maior número de regiões verificadas para formar o conjunto de dados final (COLLATZ *et al*., 2020).

É importante mencionar que todo esse procedimento é necessário, para evitar a super-representação nos dados, o que pode influenciar diretamente os algoritmos de aprendizado de máquina.

5.5 – Estratégia de Predição/Extração de Atributos

A estratégia de predição e de extração de atributos deste trabalho foi inspirada na ferramenta BEEPro (LIN *et al*.,2013), na qual apresentou resultados promissores. O objetivo da ferramenta é predizer epítopos lineares e conformacionais de células B, utilizando SVM em 16 propriedades, sendo elas: matrizes de pontuação de posição específica (PSSM), escala de proporção de aminoácidos e um conjunto de 14 propriedades físico-químicas, obtidas da base de dados do AAindex[11].

Como os cálculos da PSSM para proteomas muito grandes são demorados, inviabiliza disponibilizar uma versão *web,* desta ferramenta, para a predição de epítopos lineares de célula B. Portanto, este trabalho optou por não utilizar essa propriedade e utilizar apenas as demais supracitadas.

Diversos trabalhos da literatura (WEE *et al*., 2010; YAO *et al*., 2012) evidenciaram que determinados aminoácidos (di ou tri-peptídeos) acontecem com maior ou menor frequência em epítopos. Segundo Wee *et al*. (2010), os aminoácidos triptofano, prolina e glutamina são encontrados com maior frequência em epítopos positivos, enquanto que fenilalanina e leucina são encontrados com menor frequência. Para Yao et.al. (2012), aminoácidos como glutamina e prolina, desempenham um papel fundamental na identificação de epítopos, pois aparecem com maior frequência nos tri-peptídeos.

Diante a este fato, a taxa de proporção foi calculada para cada um dos 20 aminoácidos presentes na base de dados, por meio da equação (18). No entanto, primeiramente, é necessário calcular a frequência em que cada aminoácido ocorre em epítopos positivos e em epítopos negativos (LIN *et al*., 2013). Para tanto, a equação (1) foi aplicada individualmente em cada um dos subconjuntos (epítopos positivos e negativos) de cada conjunto de dados (DB_bac, DB_vir, DB_prot e DB_all).

[11] AAindex é uma base de dados de índices numéricos que representam várias propriedades físico-químicas e bioquímicas de aminoácidos e pares de aminoácidos (KAWASHIMA, KANEHISA 2000, KAWASHIMA *ET AL*. 2008).

$$p_{\alpha i} = \frac{f_{\alpha i}^{+} / \sum_i f_{\alpha i}^{+}}{f_{\alpha i}^{-} / \sum_i f_{\alpha i}^{-}} \tag{18}$$

sendo:

$f_{\alpha i}^{+}$: frequência do aminoácido α_i em epítopos positivos da proteína;

$f_{\alpha i}^{-}$: frequência do aminoácido α_i em epítopos negativos da proteína.

Para atenuar a dominância da frequência de alguns aminoácidos e não permitir que os algoritmos de aprendizado de máquina sejam tendenciados, a equação (19) (LOPES, 2015) os normaliza em um intervalo entre [0, 1].

$$p_{\alpha i} = \left(\frac{p_{\alpha i} - min(p_{\alpha i})}{max(p_{\alpha i}) - min(p_{\alpha i})}\right) \tag{19}$$

sendo:

$max(p_{\alpha i})$: maior valor de frequência entre os aminoácidos;

$min(p_{\alpha i})$: menor valor de frequência entre os aminoácidos.

Após testarem a combinação de várias propriedades físico-químicas e bioquímicas na predição de epítopos lineares de células B, (LIN *et al*., 2013) constataram que 14 propriedades se sobressaíram em relação as demais: hidrofilicidade (PARJ860101)[12], hidrofobicidade (PONP930101), flexibilidade (KARP850102 e BHAR880101), interatividade (BASU050101), composição (GRAR740101), volume (GRAR740103), transferência de carga (CHAM830107), capacidade de doar transferência de carga (CHAM830108), capacidade de doar ligações de hidrogênio (FAUJ880109), frequência da estrutura alfa-hélice (NAGK730101), frequência da estrutura beta (NAGK730102), frequência da estrutura coil (NAGK730103) e antigenicidade (WELLING *et al*., 1985). Tal fato justificou a utilização dessas mesmas propriedades no trabalho atual.

Obtido os valores dessas propriedades para cada aminoácido e sua respectiva taxa de proporção, é preciso varrer os epítopos de interesse, dentro da sequência completa do antígeno, para gerar os atributos que foram apresentados ao algoritmo de aprendizado de máquina. Para

[12] Códigos de cada propriedade físico-químicas e bioquímicas representados na base AAindex.

essa finalidade foi utilizado o método da janela deslizante, que utiliza uma janela móvel, atribuindo uma média ao aminoácido central da janela, na propriedade j, para j = 0, 1, ..., 14, segundo a equação 20 (LIN *et al*., 2013).

$$mediaEscala_j = \frac{\sum_i (1 - f * |c - i|) * S_i}{w} \tag{20}$$

sendo:

i: índice da posição do resíduo na janela deslizante;

c: índice da posição do resíduo central da janela;

$|c - i|$: distância em número de resíduos entre o resíduo i e o resíduo central c;

f: fator de peso linear (valor atribuído);

S_i: valor da propriedade físico-química ou taxa de proporção de aminoácido do resíduo na posição *i*.

Para definir o tamanho da janela deslizante foi observada a média do comprimento dos epítopos positivos e negativos de cada táxon e a média geral, conforme apresenta a Tabela 6. A Tabela 6 também exibe a respectiva variância.

Tabela 6 - Média e Variância do comprimento de Epítopos positivos/negativos

Táxon	Epítopos Positivos - Média	Epítopos Positivos – Variância	Epítopos Negativos – Média	Epítopos Negativos – Variância
Bactéria	13,73	27,35	9,56	0,25
Vírus	15,61	34,50	15,22	24,12
Protozoário	15,49	17,96	14,89	2,84
Total	15,26	27,18	13,89	16,24

Fonte: Próprio Autor

Com base nas médias, testes empíricos foram aplicados e tamanhos de 10, 12, 15, 17 e 20 foram testados para a janela deslizante. Ao final, a janela deslizante com tamanho igual a 20 foi a que apresentou os melhores e, portanto, a utilizada pelo trabalho.

O valor atribuído ao fator de peso linear (f) foi o valor de 0,08 recomendado pelo trabalho de (LIN *et al*., 2013). A finalidade desse fator é de aumentar a importância dos

aminoácidos vizinhos, em relação ao aminoácido em análise (o central), à medida que eles estejam mais próximos.

Após aplicar todas essas operações supracitadas, resulta-se no seguinte formato de dados:

$$< classe, valor_0, valor_1, \cdots, valor_{m-1}, valor_m >$$

sendo:

<>: uma instância utilizada no processo de treinamento;

classe: um inteiro que indica a classe que a instância pertence, 1 para epítopos positivos e 0 para epítopos negativos;

m: quantidade de propriedades;

$valor_i$: um valor real, que representa uma propriedade físico-química ou a taxa de proporção do aminoácido i.

Para evitar que atributos com valores numéricos muito grandes preponderem os valores bem pequenos, todos os valores dos atributos foram normalizados entre [0, 1].

É importante salientar que cada instância representa os cálculos das 15 propriedades aplicadas sobre um aminoácido de interesse, podendo ser tanto para o subconjunto de epítopos positivos quanto para o subconjunto de epítopos negativos.

Após os procedimentos de pré-processamento e de geração dos *embeddings*, aplicou-se a técnica de Validação Cruzada, com *K-folds* igual a 5, para dividir cada conjunto de dados em k subconjuntos mutuamente exclusivos de mesmo tamanho e, a partir de então, utilizar um subconjunto k para teste e os demais subconjuntos (k-1) para treinamento (KOHAVI, 1995).

5.6 – Proposta da Rede Neural Artificial ARTMAP-FUZZY

A RNA ARTMAP-FUZZY tem sido utilizada com bastante frequência em trabalhos de predição, principalmente no campo de engenharia elétrica, por causa de sua plasticidade, estabilidade e capacidade de convergir em poucas épocas (LOPES *et al*., 2005), (BERNARDES *et al*., 2021).

Apesar da RNA ARTMAP-FUZZY ser uma técnica que apresenta boas taxas de generalização, o desempenho de sua predição depende das escolhas dos valores dos parâmetros de vigilância: $\rho_a, \rho_b \; e \; \rho_{ab}$. Uma escolha inadequada pode resultar em perda de acurácia dos resultados, pois valores próximos de zero, permitem que os padrões poucos idênticos sejam agrupados na mesma categoria e valores próximos de um, permitem que pequenas variações

nos padrões de entrada, leve a RNA a criar novas classes (CARPENTER; GROSSBERG, 1987a).

Nesta pesquisa foram utilizados os valores de 0.61 (ρa_{baseline}), 0.8 e 0.99 para os parâmetros $\rho_a, \rho_b \; e \; \rho_{ab}$, respectivamente. O valor do parâmetro ρ_a foi definido em 1/1,618 (número áureo), visto que é uma heurística que tem dado certo. Os valores dos demais parâmetros, $\rho_b \; e \; \rho_{ab}$, foram definidos por meio de testes empíricos, bem como os valores de α e da taxa de incremento do ρ_{ab}, ambos fixados no valor de 0,1. Para o parâmetro β foi atribuído o valor de 1, pois optou-se pelo treinamento rápido, com convergência em apenas 1 época de treinamento.

As matrizes pesos $\boldsymbol{w_a}, \boldsymbol{w_b} \; e \; \boldsymbol{w_{ab}}$ foram iniciadas com os valores iguais a 1, indicando que todas as atividades estavam inativas. Inicialmente as matrizes possuem apenas 1 linha, indicando que todo início de aprendizado há apenas 1 neurônio ativo em cada matriz. À medida que ocorre o processo de treinamento e as atividades começam a ativar, novos neurônios são criados e inicializados dinamicamente. Vale mencionar que cada matriz peso deve seguir as dimensões definidas pela equação (21). Ademais, a quantidade de neurônios deve ser menor ou igual aos valores dos parâmetros $m_a, m_b \; e \; n$.

$$\begin{aligned} \boldsymbol{w_a} &= (n \; X \; m_a) \\ \boldsymbol{w_b} &= (n \; X \; m_b) \\ \boldsymbol{w_{ab}} &= (n \; X \; n) \end{aligned} \tag{21}$$

sendo:

m_a: número de componentes (atributos) dos vetores de entrada;

m_b: número de componentes do vetor de saída;

n: quantidade de padrões de entrada.

É importante ressaltar que os complementos dos padrões de entrada e de saída também são levados em consideração na definição dos valores dos parâmetros $m_a \; e \; m_b$, e, portanto, os valores desses parâmetros devem ser dobrados.

Neste estudo, como já mencionado, foi utilizado 15 propriedades para predizer se um aminoácido de uma proteína de um antígeno é um epítopo (1) ou se ele não é um epítopo (0). Desta forma, os valores dos parâmetros m_a e m_b foram definidos em 30 e 2, respectivamente. O valor de n foi definido com os valores de 35157, 97001, 109120 e 241278 para os conjuntos

de dados DB_bac, DB_prot, DB_vir e DB_all, respectivamente, pois foram essas as quantidades de *embeddings* gerados para cada conjunto de dados.

Por se tratar de um problema de classificação binária, a matriz peso w_b foi otimizada e definida de acordo com a equação (22).

$$\boldsymbol{w_b} = (2\ X\ m_b) \tag{22}$$

5.6.1 – Treinamento

A ordem de apresentação dos padrões de entrada e saída à RNA pode ser empregada de forma sequencial ou por ordem aleatória/pseudoaleatória. No entanto, do ponto de vista mental, é mais sensato adotar a ordem aleatória/pseudoaleatória. Portanto, neste trabalho foi adotada a segunda abordagem.

Para calcular o complemento dos pares de matrizes (entrada e saída) foi utilizada a estratégia de dobrar o número de linhas das respectivas matrizes ao invés de dobrar a quantidade de colunas. Desta forma, se a linha zero da matriz peso w_a contém o primeiro padrão de entrada, a linha um contém o seu complemento. Esse processo se repete para todos os demais padrões de entrada e saída, e ao final, as linhas pares conterão os padrões de entrada ou saída e as linhas ímpares os seus respectivos complementos.

Com a utilização do parâmetro de treinamento $\beta = 1$ (treinamento rápido) cada modelo de conhecimento foi gerado a partir de uma única época. Para melhorar a predição dos epítopos, para cada *fold* criado pela validação cruzada, é gerado três modelos de conhecimento, sendo que para o treinamento de cada um, os padrões de entrada e saída são apresentados a RNA de forma pseudoaleatória e em diferente ordem, possibilitando que cada modelo aprenda de um jeito e gere quantidades diferentes de neurônios.

A Tabela 7 apresenta as quantidades de neurônios que foram criados à cada modelo de conhecimento para cada partição de cada conjunto de dados, após o processo de treinamento e de validação cruzada. Vale ressaltar, que essa quantidade é apenas para as matrizes peso w_a e w_{ab}, visto que a matriz peso w_b possui apenas dois neurônios (problema de classificação binária, como já discutido anteriormente).

Tabela 7 - Número de Neurônio dos modelos de conhecimento da base de dados DB_bac, DB_vir, DB_prot e DB_all

Validação Cruzada	Modelos de Conhecimento	Quantidade de Neurônios DB_bac	Quantidade de Neurônios DB_vir	Quantidade de Neurônios DB_prot	Quantidade de Neurônios DB_all
Partição 0	Modelo 0	210	1279	165	2256
	Modelo 1	216	1315	160	2250
	Modelo 2	209	1292	171	2270
Partição 1	Modelo 0	201	1340	193	2121
	Modelo 1	194	1265	205	2157
	Modelo 2	210	1321	186	2101
Partição 2	Modelo 0	203	1315	152	1845
	Modelo 1	189	1332	193	1816
	Modelo 2	216	1363	184	1801
Partição 3	Modelo 0	208	1237	170	1826
	Modelo 1	192	1258	171	1880
	Modelo 2	203	1295	191	1812
Partição 4	Modelo 0	198	1252	183	2044
	Modelo 1	188	1307	149	1998
	Modelo 2	199	1282	170	1943

Fonte: Próprio Autor

Para gerar cada grupo de modelos de conhecimento de cada partição de cada conjunto de dados, foram utilizados as mesmas quantidades e os mesmos dados de entrada, mudando apenas a ordem em que as entradas são apresentadas à RNA.

5.6.2 – Diagnóstico

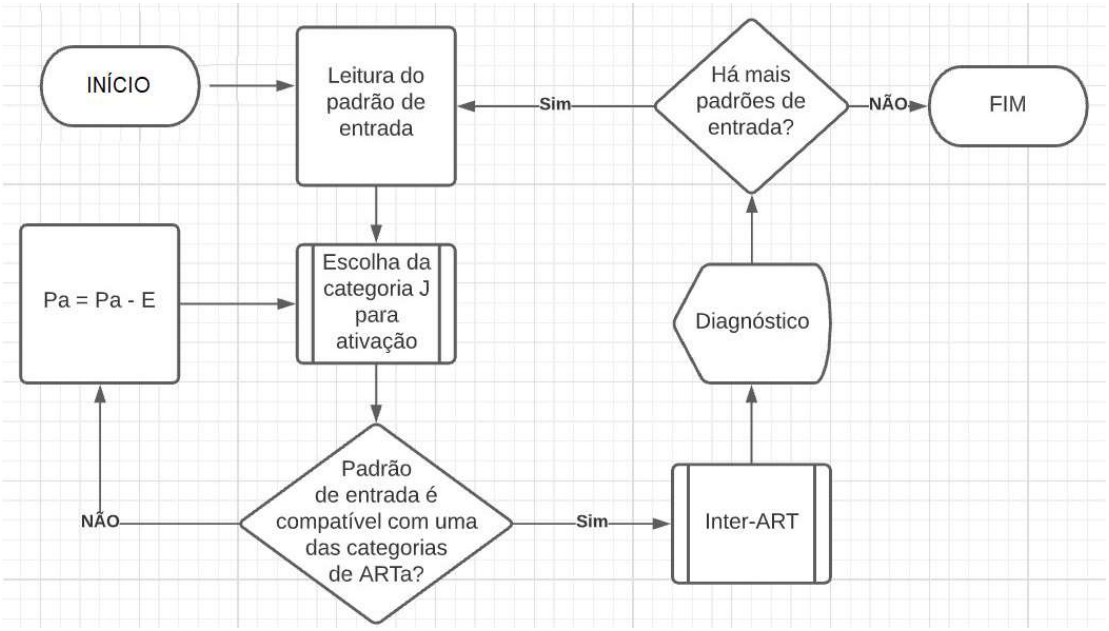
No diagnóstico a RNA ativa a categoria que melhor representa o padrão de entrada, por meio da função escolha e verifica se o grau de similaridade entre o padrão de entrada e a categoria ativa atende ao valor definido no parâmetro ρ_{a}. No entanto, é possível que apareçam padrões de entrada com características muito diferentes das categorias existentes, de forma que nenhuma das categorias seja similar o bastante para atender ao teste de vigilância.

Para esses casos foi adotada a estratégia do decremento gradual automático do parâmetro ρ_a pelo parâmetro ε, até que a categoria mais próxima se torne compatível o suficiente e passe no teste de vigilância. De forma empírica, o valor de ε foi definido em 0,1.

Após passar pelo teste de vigilância, o valor do parâmetro ρ_a volta ao seu valor original de 0.61 (ρa_{baseline}) e a execução passa ao módulo Inter-ART. Nesse instante é verificado qual a coluna da matriz peso $\boldsymbol{w_{ab}}$ está ativa, baseado no índice do neurônio vencedor da matriz peso $\boldsymbol{w_a}$ (nessa mesma linha em $\boldsymbol{w_{ab}}$). De posse do índice da coluna $\boldsymbol{w_{ab}}$, verifica-se qual a categoria (linha) que pertence a esse índice na matriz peso $\boldsymbol{w_b}$, sendo essa categoria o diagnóstico final da RNA ARTMAP-FUZZY.

A Figura 17 apresenta, simplificadamente, o processo de diagnóstico da RNA ARTMAP-FUZZY implementado.

Figura 17 - Fluxograma do processo de Diagnóstico da RNA ARTMAP-FUZZY



Fonte: Próprio Autor

Para o processo de diagnóstico foi utilizado validação cruzada e a estratégia de competição, na qual cada modelo de conhecimento gerado (3) realiza o seu próprio processo, e ao final, há competição entre os seus resultados, de forma que o resultado que prevaleça, seja o diagnóstico final da RNA, sobre o padrão de entrada corrente. Vale ressaltar que essa abordagem foi aplicada a cada conjunto de dados de maneira independente e apenas os padrões de entrada foram apresentados à RNA, os de saída foram utilizados apenas para validar os resultados de predição.

Para o conjunto de dados de teste foi aplicado apenas o processo de diagnóstico, na qual utilizou apenas os modelos de conhecimento gerados (15) a partir do conjunto de dados de treinamento DB_all. Perante a esses modelos é aplicado a técnica de competição para assim, predizer cada padrão de entrada.

5.7 – Medidas de Avaliação

Para medir o desempenho dos modelos de conhecimento gerados para cada conjunto de dados, foram utilizadas as métricas comumente empregadas na avaliação da qualidade de ferramentas de predição: sensibilidade, especificidade, precisão, acurácia e MCC, nas quais são definidas, respectivamente, pelas equações (23), (24), (25), (26) e (27) (LOPES, 2015), (LASKO *et al*., 2005).

$$R_{sen} = \frac{TP}{(TP + FN)} \tag{23}$$

$$R_{esp} = \frac{TN}{(TN + TF)} \tag{24}$$

$$R_{pres} = \frac{TP}{(TP + FP)} \tag{25}$$

$$R_{acc} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{26}$$

$$R_{mcc} = \frac{(TP * TN - FP - FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{27}$$

sendo:

TP: quantidade de verdadeiros positivos. Resíduos preditos como epítopos e realmente são epítopos;

TN: quantidade de verdadeiros negativos. Resíduos preditos como epítopos negativos e realmente são epítopos negativos;

FP: quantidade de falsos positivos. Resíduos preditos como epítopos positivos são epítopos negativos;

FN: quantidade de falsos negativos. Resíduos preditos como epítopos negativos e são epítopos positivos;

R_{sen}: a taxa de epítopos positivos que são corretamente preditas como epítopos positivos;

R_{esp}: a taxa de epítopos negativos que são corretamente preditas como epítopos negativos;

R_{pres}: dentre todas as classificações de epítopo, quantas estão corretas. Ideal em situações em que os FP são considerados mais prejudiciais do que os FN;

R_{acc}: indica um desempenho geral. O quanto o modelo classificou corretamente;

R_{mcc}: é uma medida considerada balanceada e foi proposta por Matthews em 1975. Assume valores entre -1 e 1, onde:

- Valor próximo de -1, corresponde a uma predição péssima;
- Valor próximo de 0, corresponde a uma predição aleatória; e
- Valor próximo de +1, corresponde a uma predição excelente.

Além das métricas mencionadas é utilizado um método gráfico, capaz de demostrar a relação entre a sensibilidade e a especificidade, chamado curvas ROC (do inglês, *Receiver Operating Characteristics*). No entanto, para viabilizar a comparação entre os resultados dos classificadores é preciso reduzir a curva ROC em um simples escalar. O método usualmente utilizado para esse fim é chamado AUC, cuja função é calcular a área sob a curva ROC.

5.8 – Avaliando Abordagens de Predição de Epítopos no Conjunto de Dados Teste

Para o conjunto de dados de teste, o desempenho da ferramenta proposta é comparado com a ferramenta EpiDope (COLLATZ, 2020) e com mais duas ferramentas frequentemente utilizadas na predição de epítopos lineares de células B, em suas versões mais recentes e disponíveis no IEDB: BepiPred-2.0 (JESPERSEN, 2017) e Chou e Fasman Beta Turn Prediction (WEE *et al*., 2010).

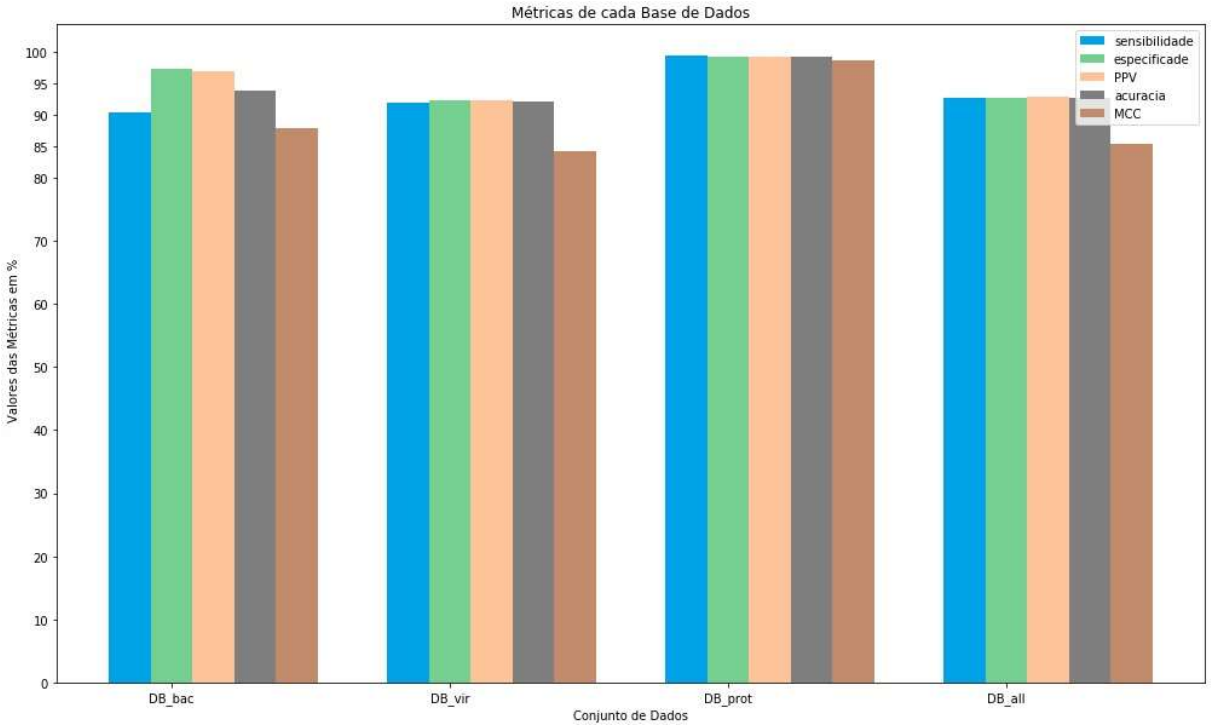
Embora sejam calculadas todas as métricas citadas na seção 3.7, na comparação das ferramentas é utilizado apenas as curvas ROC e *Precision-Recall*, pois foram as utilizadas pelo trabalho EpiDope (utilizei os seus resultados, devido à dificuldade em encontrar e utilizar as ferramentas supracitadas). Segundo Davis e Goadrich (2006), as curvas ROC podem apresentar uma visão excessivamente otimista do desempenho de um algoritmo, caso haja uma distorção na distribuição de classes. Uma alternativa bastante utilizada e eficiente nessas situações são as curvas *Precision-Recall.* As curvas ROC e as curvas *Precision-Recall* são calculadas utilizando a biblioteca *scikit-learn* (PEDREGOSA *et al*., 2011).

6 – RESULTADOS

6.1 – Conjunto de Dados de Treinamento e Validação

Após o processo de diagnóstico e de posse dos resultados das bases de dados DB_bac, DB_vir, DB_prot e DB_all, as métricas, discutidas na seção 5.7, foram aplicadas a cada resultado. A Figura 18 apresenta o panorama da relação métrica/conjunto de dados.

Figura 18 - Panorama Geral dos Resultados das Métricas de Sensibilidade, Especificidade, PPV, Acurácia e MCC sobre cada Conjunto de Dados de Teste: DB_bac, DB_vir, DB_prot e DB_all



Fonte: Próprio Autor

De acordo com a Figura 18, os resultados de todas as métricas ficaram acima de 83%, exceto o valor do MCC para a base de dados DB_vir, que foi de 70,94%. A Tabela 8 apresenta, em detalhes, o resultado de cada métrica sobre cada conjunto de dados.

Tabela 8 - Resultados Métricas/Conjunto de Dados

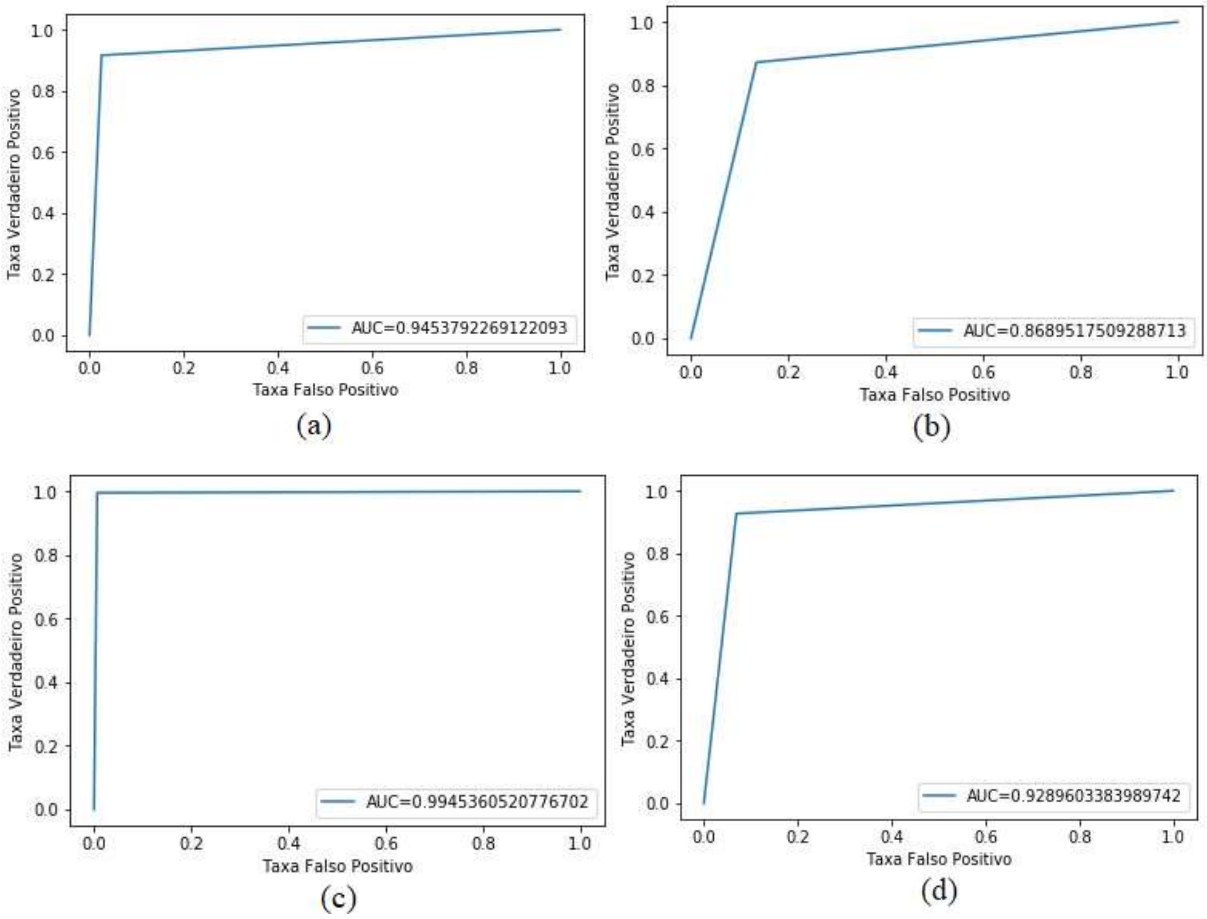
Conjunto de dados	**Métricas**				
	Sensibilidade	**Especificidade**	**PPV**	**Acurácia**	**MCC**
DB_bac	90,44%	97,20%	96,98%	93,82%	0,8788

DB_vir	83,42%	87,39%	86,88%	85,40%	0,7094
DB_prot	99,36%	99,18%	99,21%	99,27%	0,9855
DB_all	91,50%	91,49%	91,66%	91,49%	0,8300

Fonte: Próprio Autor

Buscando avaliar a capacidade do método implementado em identificar epítopos positivos, uma média da curva ROC é calculada para cada subconjunto de cada base de dados. Isso foi necessário, pois é utilizado validação cruzada quíntupla e, portanto, para cada partição é calculado uma curva ROC, e ao final, é gerado uma média da curva ROC. A Figura 19 apresenta a média das curvas ROC das bases de dados DB_bac (a), DB_vir (b), DB_prot (c) e DB_all (d). O Apêndice A apresenta a curva ROC de cada parte da validação cruzada sobre cada conjunto de dados.

Figura 19 - Área sob a curva ROC utilizando validação cruzada de 5 vezes para os Conjuntos de Dados: DB_bac (a), DB_vir (b), DB_prot (c) e DB_all (d)



Fonte: Próprio Autor

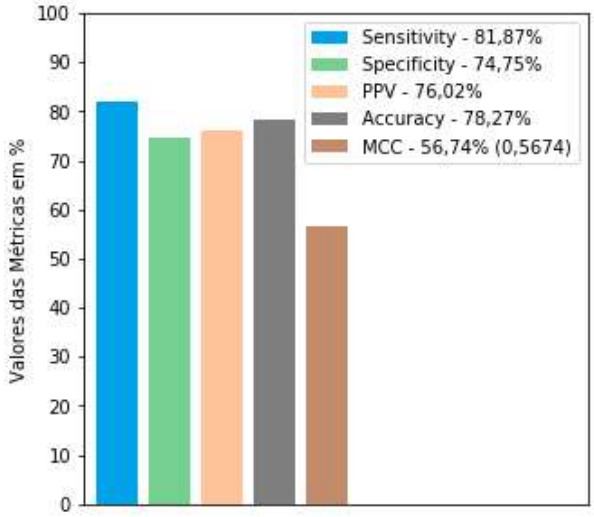
Todos os conjuntos utilizados apresentaram uma curva ROC com acurácia superior a 86%, sendo que a base de dados de protozoário (DB_prot), apresentou o melhor índice de acurácia, de aproximadamente 99,45%. O pior índice foi revelado pela base de dados viral, um valor de aproximadamente 86,89%. Os demais conjuntos ficaram dentro dessa faixa, obtendo aproximadamente os valores de 94,54% e 92,89% para DB_bac e DB_all, respectivamente.

6.2 – Conjunto de Dados de Teste

O conjunto de dados de teste foi utilizado única e exclusivamente para testar os modelos de conhecimento gerados pelo conjunto de dados DB_all. Dessa forma, não há necessidade de combinar várias previsões para gerar uma média da curva ROC, uma vez que a técnica de validação cruzada não é aplicada e os dados não são usados para treinamento. Nessa etapa foi utilizado os 15 modelos de conhecimentos, juntamente com a técnica de competição, para predizer os epítopos lineares de células B.

A partir do diagnóstico da RNA, todas as métricas brevemente discutidas na seção 5.7 são aplicadas. A Figura 20 apresenta, em detalhes, os valores alcançados.

Figura 20 - Resultados das Métricas de Sensibilidade, Especificidade, PPV, Acurácia e MCC sobre o Banco de Dados de Validação usando o modelo DB_all



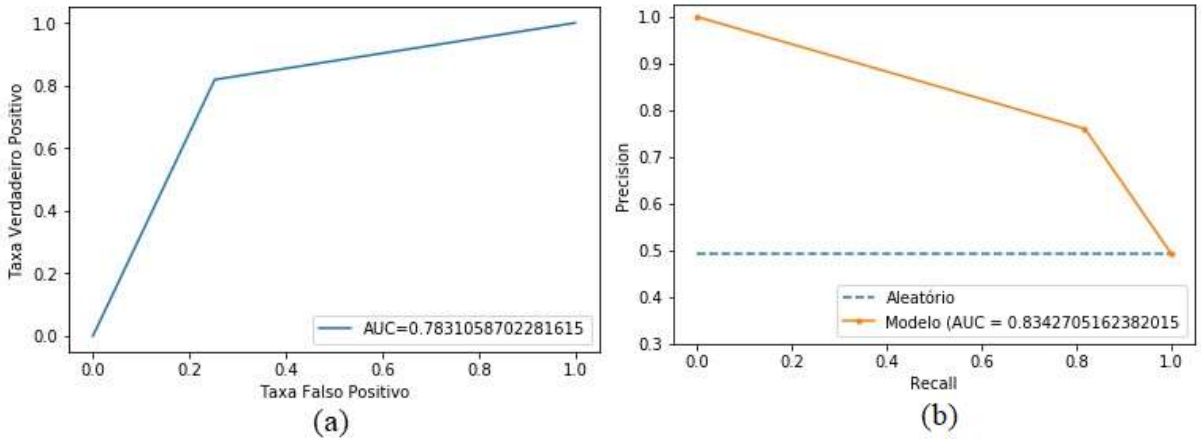
Fonte: Próprio Autor

De acordo com a Figura 20, o modelo obteve índices promissores para todas as métricas utilizadas (JESPERSEN *et al.*, 2017) (COLLATZ *et al.*, 2020) (EL-MANZALAWY; HONAVAR, 2010). O MCC com o valor de 0,5674, indica uma predição em média, superior a uma aleatória. As demais métricas obtiveram valores muito próximos, entre 74 a 82%.

Segundo Davis e Goadrich (2006) as curvas ROC podem apresentar uma visão excessivamente otimista do desempenho de um algoritmo, caso exista uma distorção na distribuição de classes (como visto na Tabela 5). Uma alternativa bastante utilizada e eficiente nessas situações são as curvas *Precision-Recall.*

A Figura 21 apresenta a curva ROC (a) e a *Precision-Recall* (b), nas quais atingiram valores de aproximadamente 0,7831 e 0,8343, respectivamente.

Figura 21 - Área sob a curva ROC (a) e Precision-Recall (b)



Fonte: Próprio Autor

6.3 – Comparação de Resultados

As principais ferramentas disponíveis na literatura não permitem testes em larga escala e limitam o acesso a muitos recursos. Desta forma, os resultados da referência (COLLATZ et. al., 2020) são utilizados para fins de comparação.

Em seu estudo, a AUC da curva ROC e da curva *Precision-Recall* são usadas para comparar o desempenho de sua ferramenta, denominada EpiDope, com o desempenho de outras ferramentas comumente utilizadas na predição de epítopos lineares de células B.

A partir dos resultados obtidos com o banco de dados de teste, o mesmo conjunto utilizado por (COLLATZ et. al., 2020) e que passou pelos mesmos procedimentos de pré-processamentos, buscou-se compará-los com os resultados das demais ferramentas.

A Tabelas 9 apresenta os resultados da AUC da curva ROC e da curva *Precision-Recall*, da ferramenta BepFANM, das ferramentas mencionadas e de uma predição média aleatória. Nota-se que a ferramenta BepFANM (proposta) atingiu o valor de aproximadamente 0,7831 para a AUC da curva ROC e 0,8343 para a AUC da curva *Precision-Recall*, superando os valores das demais ferramentas, inclusive de uma predição média aleatória.

Tabela 9 - Comparação da AUC da curva ROC e da curva *Precision-Recall* entre as ferramentas (Adaptado de COLLATZ *et al*., 2020)

Ferramenta	AUC - ROC	AUC – *Precision Recall*
BepFAMN	0,783	0,83
EpiDope	0,605	0,66
BepiPred-2.0	0,465	0,56
Betaturn	0,562	0,64
Aleatório	0,500	0,58

Fonte: Próprio Autor

7 – DISCUSSÃO

Existem várias ferramentas computacionais disponíveis na literatura para a identificação *in sílico* de epítopos lineares de células B. Conforme descrito na seção 3, técnicas de aprendizado de máquina como SVM, RNA e Modelos Ocultos Markovianos, por exemplo, combinadas com outra abordagem, geralmente com o método de escala de propensão ou outros métodos híbridos (LIN *et al*., 2013), são comumente utilizadas para esse fim.

No entanto, muitas dessas ferramentas apresentam baixo índice de precisão na predição de epítopos, ou quando melhoram os resultados, envolvem altos custos computacionais, inviabilizando o uso em larga escala, principalmente via Internet.

Há diversos fatores que corroboram com essa baixa taxa de acerto, sendo o conjunto de dados utilizado para o treinamento um dos principais. Geralmente, as bases de treinamento são muito redundantes e apresentam um desequilíbrio grotesco na distribuição dos dados entre as classes. Fato que torna indispensável a aplicação de um bom pré-processamento para mitigar tais problemas.

A base de dados utilizada neste estudo foi extraída do site IEDB e passou por diversas etapas de pré-processamento para minimizar os problemas citados e gerar os *embeddings* que serão apresentados ao algoritmo de aprendizado de máquina para extração de correlações. No entanto, o conjunto de dados final foi bastante reduzido, o que pode afetar diretamente os resultados dos modelos de conhecimento, uma vez que dados importantes podem ser eliminados nesse processo.

Um exemplo prático desse problema aconteceu com a base de dados DB_prot, formada por dados de protozoários. Inicialmente esse conjunto de dados continha um grande desequilíbrio entre as classes, haviam 4137 epítopos positivos contra 17275 epítopos negativos. Após a etapa de pré-processamento as classes ficaram mais balanceadas, reduzindo as suas quantidades para 3992 e 3991, respectivamente.

Observe que o número de epítopos positivos também diminui ligeiramente. Isso acontece, porque existem muitos epítopos semelhantes, o que poderia prejudicar a capacidade de generalização do algoritmo de aprendizado e fazer com que ele simplesmente memorize os dados. Assim, é muito importante aplicar técnicas/ferramentas auxiliares que agrupem epítopos semelhantes, com base em um limiar, de modo que, ao final, apenas um epítopo por grupo forme o conjunto de dados. Um ponto interessante a ser discutido é o valor desse limiar, ou seja, o quanto os epítopos positivos/negativos devem ser diferentes a ponto de não permitirem que os algoritmos de aprendizado de máquina os memorizem. Neste estudo, foi utilizado a

recomendação de (GAO *et al*., 2012), para o conjunto de dados de treinamento e validação. Para o conjunto de dados de teste foi utilizado o limiar recomendado por (COLLATZ, 2020).

O método implementado para a predição de epítopos foi inspirado nos trabalhos (LIN *et al*., 2013) e (LOPES, 2015) e utiliza 14 propriedades físico-químicas dos aminoácidos e a sua proporção em epítopos positivos e negativos. Optou-se por não utilizar a informação evolutiva, codificada como PSSM (LIN *et al*., 2013), devido ao seu alto custo computacional, impossibilitando qualquer futura possibilidade de disponibilização da ferramenta remotamente. Mesmo sem a PSSM, a RNA ARTMAP-FUZZY apresentou resultados promissores, tanto para a base de dados de treinamento e validação quanto para a base de teste.

A RNA ARTMAP-FUZZY foi treinada e testada em diferentes conjuntos de dados, cada um formado por um único táxon e um formado pela união dos 3 táxons. Para cada conjunto foi aplicada a validação cruzada de 5 vezes e a técnica de competição de 3 vezes, e, portanto, foram gerados 3 modelos de conhecimento para cada execução da validação cruzada, totalizando em 15 modelos para cada conjunto de dados. Desta forma, a cada 3 modelos gerados é feito o diagnóstico e calculado as métricas, e ao final do processo de validação cruzada, uma média para cada métrica é calculada.

A técnica de competição acrescentou cerca de 5 a 15% ao valor final de cada métrica. Isso se justifica, pois a rede cria modelos independentes que aprendem de maneiras diferentes, para o mesmo conjunto de dado, em cada execução. A ordem em que os padrões de entrada alimentam a rede e o valor de seus parâmetros influenciam como os modelos aprendem. Assim, para cada particionamento construído pela validação cruzada sobre cada conjunto de dados, o algoritmo é executado três vezes e, ao final, o resultado que ocorrer com maior frequência, para cada padrão de entrada, é a predição da RNA.

A RNA ARTMAP-FUZZY utilizou aprendizagem rápida na etapa de treinamento, usando apenas uma época para cada modelo de conhecimento gerado. É claro que essa característica é uma das grandes vantagens dos algoritmos da família ART em relação a outras abordagens disponíveis na literatura, que geralmente utilizam várias épocas para convergirem.

Os testes realizados no conjunto de dados DB_prot mostraram as classificações de desempenho mais altas, todas acima de 99%. No entanto, a base de dados viral (DB_vir), mesmo com boas taxas, apresentou os valores mais baixos para todas as métricas. Esse fato indica que a RNA ARTMAP-FUZZY teve uma dificuldade maior para a generalização deste táxon, visto que a quantidade de neurônios criados foi muito maior (Tabela 7) comparada a quantidade criada para os modelos dos demais táxons. Outro indício que corrobora com esse

resultado é o fato de o valor da variância ser consideravelmente superior ao valor das demais, tanto para os epítopos positivos quanto para os negativos (Tabela 6).

O MCC é uma métrica de qualidade utilizada em técnicas de aprendizado de máquina para classificação binária. Descreve se o resultado do algoritmo representa uma perfeita predição (valores próximos de 1), uma predição aleatória (valores próximos de 0) ou uma predição ruim (valores próximos de -1). Neste sentido, entende-se que o resultado do MCC para todos os táxons foi satisfatório, uma vez que os valores alcançados foram todos superiores a 0,70. Este fato demonstra a boa capacidade de generalização que a RNA ARTMAP-FUZZY obteve para esse tipo de dados.

O conjunto de dados DB_all, o que contém dados de todos os táxons, foi utilizado para gerar os modelos de conhecimento que foram aplicados ao conjunto de dados de validação e ao conjunto de teste. No conjunto de validação, ele obteve classificações de alto desempenho, sendo ligeiramente melhor, em todas as métricas, do que o conjunto de dados DB_vir.

No conjunto de dados de teste os modelos de conhecimento obtiveram índices aceitáveis de desempenho, alcançando uma taxa de sensibilidade de 81,84%, mesmo com a quantidade de resíduos de epítopos negativos muito superior a quantidade de epítopos positivos. Portanto, dizer deterministicamente que todos os resíduos são epítopos negativos pode fornecer uma alta taxa de especificidade, mas não de sensibilidade. Por outro lado, uma especificidade de 74,75% também garante que o modelo simplesmente não preveu todos os resíduos como epítopos positivos, caso contrário teria uma especificidade muito baixa.

Devido as características apresentadas pelo conjunto de dados de teste, um MCC com valor de aproximadamente 0,5674, pode ser interpretado como um índice de predição geral melhor do que um aleatório.

Na comparação com as outras ferramentas de predição, utilizou-se o conjunto de dados de teste e as métricas AUC da curva ROC e da curva *Precision-Recall*. Em ambas, a BepFAMN superou as abordagens de predição das demais ferramentas, sendo que a mais próxima, a EpiDope, atingiu valores de 0,605 e 0,66, respectivamente. Apesar de (JESPERSEN *et al*., 2017) documentar que a sua ferramenta BepiPred-2.0 atingiu um desempenho de 0,57 para a AUC da curva ROC, neste conjunto de dados reduzido e menos redundante não foi possível reproduzir o mesmo resultado, sendo inferior até mesmo a uma predição média aleatória (COLLATZ *et al*., 2020).

Com valores de aproximadamente 0,7831 e 0,8343 para a AUC da curva ROC e para AUC da curva *Precision-Recall*, respectivamente, garante que a ferramenta BepFAMN, além

de superar uma predição média aleatória, superou todas as outras ferramentas, inclusive, em ambas as curvas.

Uma das grandes vantagens das redes neurais da família ART é a sua capacidade de proporcionar aprendizado continuado, ou seja, ela pode aprender continuamente novos padrões sem esquecer o que já aprendeu, ou seja, a RNA consegue atualizar o seu modelo de conhecimento apenas com os novos dados. Esse mecanismo não acontece em outros tipos de RNAs disponíveis na literatura, nas quais precisam treinar novamente com todos os dados (os antigos mais os recentes) para gerar os seus modelos de conhecimento atualizados. Neste sentido, viabilizar uma ferramenta que utiliza um processo de aprendizagem incremental, alternando automaticamente entre os estados de aprendizado e de diagnóstico, é muito promissor.

8 – CONCLUSÃO

Esta pesquisa tem contribuído oferecendo uma nova abordagem para a resolução do problema-alvo baseada nas redes neurais artificiais (ANN) usando a teoria de ressonância adaptativa (ART). Ainda que esta arquitetura tenha sido proposta, há bastante tempo (década de 80), é, por certo, atual e bastante competitiva em relação às novas técnicas disponíveis na literatura especializada como o "deep learning". Ressalta-se que as ANN da família ART também são susceptíveis a implementação do conceito "deep learning", inclusive, trata-se de uma abordagem que está sendo desenvolvida pelo grupo de pesquisa.

Neste sentido, abre perspectiva de aplicações em várias áreas do conhecimento humano, incluindo, em especial, dar continuidade de pesquisas no contexto da saúde, por exemplo, auxiliar no desenvolvimento de imunizantes. Considera-se uma importante contribuição da área tecnológica à saúde pública buscando estabelecer uma saudável "simbiose", proporcionando uma ótima oportunidade de desenvolver novas arquiteturas de ANN.

A ferramenta apresentada neste estudo é capaz de predizer epítopos lineares de células B, exigindo apenas uma sequência de aminoácidos como entrada. A BepFAMN demonstrou o melhor desempenho dentre as ferramentas frequentemente utilizadas para esse tipo de predição.

Todas as proteínas dos conjuntos de dados de treinamento/validação possuem semelhança inferior a 80% e do conjunto de teste, inferior a 80% e 50% (foram aplicados os dois limiares, conforme o processo descrito pela Figura 16). Este fato garante a independência entre os conjuntos e não permite que a RNA ARTMAP-FUZZY memorize os dados de treinamento, mas aumente a sua capacidade de generalização.

O desempenho da BepFAMN pode ser atribuído às características de estabilidade e de plasticidade fornecidas pelas redes da família ART, pela sua boa capacidade de generalização para esse tipo de dado e pela contribuição da técnica de competição. Ademais, como não foi utilizado o PSSM em seu pré-processamento, viabiliza o desenvolvimento de uma versão Web e/ou Stand Alone para uso em larga escala, sem afetar o desempenho.

Finalmente, as ferramentas de predição de epítopos devem servir principalmente como filtros para descartar regiões improváveis de serem epítopos e, assim, eliminar análises experimentais desnecessárias. Desta forma, devem surgir ferramentas que apresentem índices aceitáveis de sensibilidade e especificidade, de forma que permitam que esses experimentos sejam mais precisos e direcionados.

8.1 – Sugestões para Trabalhos Futuros

Para dar sequência ao estudo de predição *in sílico* de epítopos lineares de células B, sugerem-se os seguintes tópicos:

- Desenvolver uma estratégia para extrair o perfil PSSM de forma eficiente, para utilizá-lo no processo de treinamento, uma vez que já foi reportado pela comunidade como uma propriedade promissora;
- Disponibilizar a ferramenta via *Web*, capaz de fornecer, automaticamente, aprendizado continuado, ou seja, de acordo com as entradas fornecidas pelo usuário, a ferramenta ativa o processo de treinamento incremental ou o processo de diagnóstico;
- Implementar melhorias de otimização na estrutura da RNA ARTMAP-FUZZY;
- Implementar outra estratégia de aprendizado, ainda não utilizado, como Sistemas Imunológicos Artificiais, e comparar os resultados

REFERÊNCIAS

ABBAS A. K.; LICHTMAN, A. H. **Cellular and molecular immunology**. 9. ed. Saunders: [s.n.], 2019.

ALTSCHUL S. F.; GISH W.; MILLER W.; MYERS E. W.; LIPMAN D. J. Basic local alignment search tool. **Journal of Molecular Biology**, London, v. 215, p. 403-410, 1990. DOI: https://doi.org/10.1016/S0022-2836(05)80360-2.

BAIROCH, A.; APWEILER, R. **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000**. Nucleic Acids Res 2000; 28: 45–48. DOI: https://doi.org/10.1093/nar/28.1.45.

BAIROCH, A.; APWEILER, R.; WU, C. H.; BARKER, W. C.; BOECKMANN, B.; FERRO, S.; GASTEIGER, E.; HUANG, H.; LOPEZ, R.; MAGRANE, M.; MARTIN, M. J.; NATALE, D. A.; O'DONOVAN, C.; REDASCHI, N.; YEH, L. S. **The universal protein resource (UniProt)**. Nucleic Acids Res. 2005 Jan 1;33 (Database issue): D154-9. DOI: https://doi.org/10.1093/nar/gki070.

BERNARDES, H. R. S.; TONELLI-NETO, M. S.; MINUSSI, C. R. Fault classification in power distribution systems using multiresolution analysis and a fuzzy-ARTMAP neural network. **IEEE Latin America Transactions**, Piscataway, v. 19, p. 1824-1831, 2021. DOI: https://doi.org/10.1109/TLA.2021.9475615.

BLYTHE, M. J.; FLOWER, D. R. Benchmarking B cell epitope prediction: Underperformance of existing methods. **Protein Science**, Hoboken, v. 14, n. 1, p. 246-248, 2005. DOI: http://dx.doi.org/10.1110/ps.041059505] [PMID: 15576553.

CALL, M. E.; WUCHERPFENNIG, K. W. Common themes in the assembly and architecture of activating immune receptors. **Nature Reviews Immunology**, London, v. 7, p. 841-850, 2007. DOI: 10.1038/nri2186.

CAMACHO C.; MADDEN T.; COULOURIS, G.; AVAGYAN, V.; MA, N.; TAO, T.; AGARWALA, R. **BLAST command line applications user manual**. 2013. DOI: http://nebc.nerc.ac.uk/bioinformatics/documentation/blast+/user_manual.pdf.

CARMONA, S. J.; SARTOR, P. A.; LEGUIZAMÓN, M. S.; CAMPETELLA, O. E.; AGÜERO, F. **Diagnostic peptide discovery**: prioritization of pathogen diagnostic markers using multiple features. **PloS One**, San Francisco, v. 7, e50748, 2012. DOI: https://doi.org/10.1371/journal.pone.0050748.

CARPENTER, G. A.; GROSSBERG, S. A massively parallel architecture for a self-organization neural pattern recognition machine. **Computer Vision, Graphics, and Image Processing**, Maryland Heights, v. 37, p. 54-115, 1987a.

CARPENTER, G. A.; GROSSBERG, S. ART2: Self-organization of stable category recognition codes for analog input patterns. **Applied Optics: Special Issue on Neural Networks**, New York, v. 26, p. 4919–4930, 1987b.

CARPENTER, G. A.; GROSSBERG, S.; REYNOLDS, J. H. ARTMAP: supervised real-learning and classification of non-stationary data by a selforganizing neural network. **Neural Network**, Marietta, v. 4, n. 5, p. 565-588, 1991a.

CARPENTER, G. A.; GROSSBERG, S.; ROSEN, D. B. Fuzzy ART: fast stable learning and categorization of analog patterns by an adaptive resonance system. **Neural Networks**, Kidlington, v. 4, n. 6, p. 759-771, 1991b.

CARPENTER, G. A. *et al*. Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. **IEEE Transactions on Neural Networks**, Piscataway, v. 3, n. 5, p. 698-713, 1992. DOI: https://doi.org/10.1109/72.159059.

CHOW, A.; BROWN B. D.; MERAD, M. Studying the mononuclear phagocyte system in the molecular age. **Nature Reviews Immunology**, London, v. 11, p. 788-798, 2011. DOI: 10.1038/nri3087.

CERUTTI, A. The regulation of IgA class switching. **Nature Reviews Immunology**, London, v. 8, p. 421 - 434, 2008. DOI: 10.1038/nri2322.

CHEN, J.; LIU, H.; YANG, J.; CHOU, K. C. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. **Amino Acids**, Wien, v. 33, p. 423-428, 2007. DOI: https://doi.org10.1007/s00726-006-0485-9.

COLLATZ, M.; MOCK, F.; HÖLZER, M.; BARTH, E.; SACHSE, K.; MARZ, M. **EpiDope**: A Deep neural network for linear B-cell epitope prediction. DOI: bioRxiv 2020.05.12.090019; doi: https://doi.org/10.1101/2020.05.12.090019.

DANILOVA, N.; AMEMIYA, C. T. Going adaptive: the saga of antibodies. **Annals of the New York Academy of Sciences**, New York, v. 11, p. 130 – 155, 2009. DOI: 10.1111/j.1749-6632.2009.04881.x.

DAVIS, J.; GOADRICH, M. The relationship between Precision-Recall and ROC curves. *In:* ICML '06, PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 23, 2006. **Proceedings of the** […]. p. 233–240. DOI: https://doi.org/10.1145/1143844.1143874.

DAVYDOV, Y. I.; TONEVITSKY, A. G. Prediction of linear B-cell epitopes. **Molecular Biology**, Moscow, v. 43, p. 150–158. 2009. DOI: https://doi.org/10.1134/S0026893309010208.

EL-MANZALAWY, Y.; HONAVAR, V. Recent advances in B-cell epitope prediction methods. **Immunome Research**, Belgium, v. 6, n. 2, 2010. DOI: https://doi.org/10.1186/1745-7580-6-S2-S2.

EL-MANZALAWY, Y.; DOBBS, D.; HONAVAR, V. Predicting linear B-cell epitopes using string kernels. **Journal Molecular Recognition**, Oxford, v. 21, p. 243-255, 2008. DOI: https://doi.org/10.1002/jmr.893.

FAGARASAN, S. Evolution, development, mechanism and function of IgA in the gut. **Current Opinion in Immunology**, Oxford, v. 20, n. 2, p. 170-177, 2008. DOI: 10.1016/j.coi.2008.04.002.

FARBER, D. L.; YUDANIN, N. A.; RESTIFO, N. P. Human memory T cells: generation, compartmentalization and homeostasis. **Nature Reviews Immunology**, London, v. 14, p. 24-35, 2014. DOI: 10.1038/nri3567.

JESPERSEN, M. C.; PETERS, B.; NIELSEN, M.; MARCATILI, P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. **Nucleic Acids Research**, Oxford, v. 45(W1), W24–W29, 2017. DOI: https://doi.org/10.1093/nar/gkx346.

GAO, J.; FARAGGI, E.; ZHOU, Y.; RUAN, J.; KURGAN L. **BEST**: Improved predictionof B-Cell epitopes from antigen sequences. **PLoS** ONE, v. 7, p. e40104. 2012.

GARCIA, E. S. O passo seguinte do genoma. **Ciência Hoje**, São Paulo, v. 24, n. 144, p. 50-51, 1998.

GILBERT, W. Rumo ao proteoma. **Ciência Hoje**, São Paulo, v. 29, n. 173, p. 8-11, 2001.

GOLDSBY, R.; KINDT, T. J.; OSBORNE, P. A.; KUBY, J. **Immunology**. 5. ed. New York: W. H. Freeman, 2003.

GOLDSCHMITD, R.; PASSOS, E. **Data mining**: um guia prático – conceitos, técnicas, ferramentas, orientações e aplicações. Rio de Janeiro: Elsevier, 2005.

GREENBAUM, J. A.; ANDERSEN, P. H.; BLYTHE, M.; BUI, H. H.; CACHAU, R. E.; CROWE, J.; DAVIES, M.; KOLASKAR, A. S.; LUND, O.; MORRISON, S.; MUMEY, B.; OFRAN, Y.; PELLEQUER, J. L.; PINILLA, C.; PONOMARENKO, J. V.; RAGHAVA, G. P.; VAN REGENMORTEL, M. H.; ROGGEN, E. L.; SETTE, A.; SCHLESSINGER, A.; SOLLNER, J.; ZAND, M.; PETERS, B. Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. **Journal Molecular Recognition**, Oxford, v. 20, n. 2, p. 75-82, 2007.

GROSSBERG, S. Adaptive resonance theory: how a brain learns to consciously attend, learn, and recognize a changing world. **Neural Networks**, Oxford, v. 37, p. 1-47, 2013. DOI: 10.1016/j.neunet.2012.09.017.

HARWOOD, N. E.; BATISTA, F. D. Early events in B cell activation. **Annual Review of Immunology**, Palo Alto, v. 28, p. 185 – 210, 2010. DOI: 10.1146/annurev-immunol-030409-101216.

HOCHREITER, S.; SCHMIDHUBER J. Long shortterm memory. **Neural Computation**, Cambridge, v. 9, n. 8, p. 1735–1780, 1997. DOI: https://doi.org/10.1162/neco.1997.9.8.1735.

HOPP T. P.; WOODS K.R. Prediction of protein antigenic determinants from amino acid sequences. **Proceeding of the National Academic Science U S A**. [s.l.], v. 78, n. 6, p. 3824-2828, 1981. DOI: https://doi.org/10.1073/ pnas.78.6.3824.

KAWASHIMA, S.; KANEHISA, M. AAindex: amino acid index database. **Nucleic Acids Research**, Oxford, v. 28, p. 374–374, 2000.

KAWASHIMA, S.; POKAROWSKI, P.; POKAROWSKA, M.; KOLINSKI, A.; KATAYAMA, T.; KANEHISA, M. AAindex: amino acid index database, progress report. **Nucleic Acids Research**, Oxford, v.36, p. D202–D205, 2008.

KINDT, T. J.; GOLDSBY, R. A.; OSBORNE, B. A. **Imunologia de kubi**. 6. ed. São Paulo: Jones & Bartlett, 2008.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *In*: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 14, 1995, Quebec. **Proceedings of the** […]. Quebec: [s.n.], 1995. p. 1137–1145.

KORBER, B.; BRANDER, C.; HAYNES, B.; KOUP, R.; MOORE, J.; WALKER, B.; WATKINS, D. **HIV immunology and HIV/SIV vaccine databases**. Los Alamos: New Mexico Los Alamos National Laboratory, Theoretical Biology and Biophysics, 2003.

KORBER, B.; LABUTE, M.; YUSIM, K. Immunoinformatics comes of age. **PloS Computational Biology**, San Francisco, v. 2, n. 6, p. 484-492, 2006.

KORB, K. B.; NICHOLSON, A. E. **Bayesian artificial intelligence**. Florida: Chapman & Hall/CRC, 2003.

KRINGELUM, J. V.; NIELSEN, M.; PADKJÆR, S. B.; LUND, O. Structural analysis of B-cell epitopes in antibody:protein complexes. **Molecular Immunology**, Oxford, v. 53, n. 1-2, p. 24-34, 2013. DOI: 10.1016/j.molimm.2012.06.001.

KUROSAKI, T.; SHINOHARA, H.; BABA, Y. B cell signaling and fate decision. **Annual Review of Immunology**, Palo Alto, v. 28, p. 21-55, 2010. DOI: 10.1146/annurev.immunol. 021908.132541.

LASKO, T. A.; BHAGWAT, J. G.; ZOU, K. H.; OHNO-MACHADO, L. The use of receiver operating characteristic curves in biomedical informatics. **Journal of Biomedical Informatics**, Maryland Heihts, v. 38, n. 5, p. 404-415, 2005. DOI https://doi.org/10.1016/j.jbi.2005.02.008.

LI, W.; GODZIK, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. **Bioinformatics**, Oxford, v. 22, n. 13, p. 1658 - 1659, 2006.

LIMA, F. A.; SAMPAIO, M. C. O papel do timo no desenvolvimento do sistema imunológico. **Pediatria**, São Paulo, v. 29, n. 1, p. 33-42, 2007.

LIN, S. Y. H.; CHENG, C. W.; SU, E. Y. Prediction of B-cell epitopes using evolutionary information and propensity scales. **BMC Bioinformatics**, London, v. 14, p. S10, 2013. DOI https://doi.org/10.1186/1471-2105-14-S2-S10.

LITMAN, G. W.; RAST, J. P.; FUGMANN, S. D. The origins of vertebrate adaptive immunity. **Nature Reviews Immunology**, London, v. 10, p. 543 – 553, 2010. DOI: 10.1038/nri2807.

LOPES, M. L. **Desenvolvimento de redes neurais para previsão de cargas elétricas de sistemas de energia elétrica**. 2005. Tese (Doutorado em Engenharia Elétrica) – Faculdade de Engenharia, Universidade Estadual Paulista, Ilha Solteira, 2005.

LOPES R. S. **Desenvolvimento de ferramentas para a identificação de marcadores moleculares e imunológicos a partir de dados genômicos como alvo para o diagnóstico de doenças parasitárias**. 2015. Tese (Doutorado) - Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, 2015.

LORENA, A. C.; CARVALHO, A. C. P. L. F. Uma introdução às support vector machine. **Revista Informática Teórica e Aplicada**, Porto Alegre, v. 14, p. 43-67, 2007. DOI: https://doi.org/10.22456/2175-2745.5690.

MACHADO DE ÁVILA, R. A. **Predição de epítopos descontínuos ou conformacionais em proteínas através da bioinformática estrutural**. 2011. 137 f. Tese (Doutorado) - Curso de Pós-Graduação em Bioinformática, Universidade Federal de Minas Gerais, Belo Horizonte, 2011.

MATTHEWS, B. W. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. **Biochimica et Biophysica Acta (BBA) - Protein Structure**, v. 405, n. 2, p. 442–451, 1975.

MAYER, M.; MEYER, B. Group epitope mapping by saturation transfer difference NMR to identify segments of a ligand in direct contact with a protein receptor. **Journal American Chemical Society**, Washington, v.123, n. 25, p. 6108-6117, 2001.

MEDZHITOV, R.; JANEWAY, C. J. R. Innate immunity. **The New England Journal Medicine**, Waltham, v. 343, p. 338-344, 2000. DOI: https://doi.org/10.1056/NEJM200008033430506.

NCBI **– National Center for Biotechnology Information**. Disponível em: http://www.ncbi.nlm.nih.gov. Acesso em: 31 mar. 2021.

NCBI – National Center for Biotechnology Information. **NIH – National library of medicine**. Disponível em: http:// pubmed.ncbi.nlm.nih.gov. Acesso em: 15 fev. 2022.

MURPHY K. **Imunobiologia de Janeway**. 8. ed. Porto Alegre: Artmed. 2014.

PARHAM P. **O sistema imune**. 3. ed. Porto Alegre: Artmed. 2011.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: machine learning in python. **Journal of Machine Learning Research**, Cambridge, v. 12, p. 2825–2830, 2011.

PELLEQUER, J.; WESTHOF, E.; VAN REGENMORTEL, M. H. V. Correlation between the location of antigenic sites and the prediction of turns in proteins. **Immunology Letters**, Amsterdam, v. 36, p. 83-99, 1993. DOI: https://doi.org/10.1016/0165-2478(93)90072-a.

PETERS, B.; SIDNEY, J.; BOURNE, P.; BUI, H. H.; BUUS, S.; DOH, G.; FLERI, W.; KRONENBERG, M.; KUBO, R.; LUND, O.; NEMAZEE, D.; PONOMARENKO, J. V.; SATHIAMURTHY, M.; SCHOENBERGER, S.; STEWART, S.; SURKO, P.; WAY, S.; WILSON, S.; SETTE, A. The immune epitope database and analysis resource: from vision to blueprint. **PLoS Biology**, San Francisco, 3. 2005. DOI: https://doi.org/10.1371/journal.pbio.0030091.

PONOMARENKO, J. V.; REGENMORTEL, M. H. V. VAN. B-Cell epitope prediction. *In*: GU, J.; BOURNE, P. (ed.). **Structural bioinforma**. [New York]: John Wiley & Sons. 2009. p. 849-879.

PONOMARENKO, J. V.; BOURNE, P. Antibody-protein interactions: benchmark datasets and prediction tools evaluation. **BMC Structural Biology**, Londpn, v. 7, 2007. DOI: https://doi.org/10.1186/1472-6807-7-6464.

RICKERT, R. C. New insights into pre-BCR and BCR signalling with relevance to B cell malignancies. **Nature Reviews Immunology**, London, v. 13, p. 578 - 591, 2013. DOI: 10.1038/nri3487.

RUX, J. J.; BURNETT, R. M. Type-specific epitope locations revealed by X-ray crystallographic study of adenovirus type 5 hexon. **Molecular Therapy**, Cambridge, v. 1, n. 1, p. 18-30, 2000. DOI: http://dx.doi.org/ 10.1006/mthe.1999.0001] [PMID: 10933908.

SAHA, S.; RAGHAVA, G. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. **Proteins**, Hoboken, v. 65, p. 40-48, 2006.

SAHA, S.; BAHSIN, M.; RAGHAVA, G. P. Bcipep: a database of B-Cell epitopes. **BMC Genomics**, London, v. 6, p. 79, 2005. DOI: https://doi.org/10.1186/1471-2164-6-79.

SANTOS JÚNIOR, C. R. **Uma nova abordagem de treinamento on-line para rede neural artificial ARTMAP-FUZZY**. 2017. Tese (Doutorado) – Faculdade de Engenharia, Universidade Estadual Paulista, Ilha Solteira, 2017.

SATHIAMURTHY, M.; PETERS, B.; BUI, H. H.; SIDNEY, J.; MOKILI, J.; WILSON, S. S.; FLERI, W.; McGUINNESS, D.; BOURNE, P. E.; SETTE, A. An Ontology for immune epitopes: application to the design of a broad scope database of imune reactivities. **Immunome Research**, v. 1, n. 2, 2005.

SHAO, J.; XU, D.; TSAI, S. N.; WANG, Y.; NGAI, S. M. Computational identification of protein methylation sites through bi-profile Bayes feature extraction. **PLoS One**, San Francisco, v. 4, n. 3, p. 4920, 2009. DOI: https://doi.org/10.1371/journal.pone.0004920.

SILVERSTEIN, A. M. Cellular versus humoral immunology: a century-long dispute. **Nature Immunology**, New York, v. 4, p. 425–428, 2003.

SOLLNER, J.; GROHMANN, R.; RAPBERGER, R.; PERCO, P.; LUKAS, A.; MAYER, B. Analysis and prediction of protective continuous B-cell epitopes on pathogen proteins. **Immunome Research**, v. 4, n. 1, 2008. DOI: https://doi.org/10.1186/1745-7580-4-1.

SUN, P.; GUO, S.; SUN, J.; TAN, L.; LU, C.; MA, Z. Advances in In-silico B-cell epitope prediction. **Current Topics in Medicinal Chemistry**, Sharjah, v. 19, p. 105, 2019. DOI: https://doi.org/10.2174/1568026619666 181130111827.

TOSELAND, C.; CLAYTON, D.; MCSPARRON, H.; HEMSLEY, S.; BLYTHE, M.; PAINE, K.; DOYTCHINOVA, I.; GUAN, P.; HATTOTUWAGAMA, C.; FLOWER, D. AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. **Immunome Research**, v. 1, n. 1, p. 4, 2005. DOI: https://doi.org/10.1186/1745-7580-1-4.

TRAVIS, J. On the origins of the immune system. **Science**, Washington, v. 324, p. 580–582, 2009. DOI: 10.1126/science.324_580. https://doi.org/10.1126/science.324_580.

VAN REGENMORTEL, M. H. **What is a B-cell epitope? Epitope mapping protocols**. [New York: Springer, 2009. DOI: https://doi.org/10.1007/978-1-59745-450-6_1.

VAN REGENMORTEL, M. H. **Antigenicity and immunogenicity of synthetic peptides**. **Biologicals**, v. 29, n.3-4, p. 209-213, 2001. DOI: 10.1006/biol.2001.0308. PMID: 11851317.

VITA, R.; MAHAJAN, S.; OVERTON, J. A.; DHANDA, S. K.; MARTINI, S.; CANTRELL, J. R.; WHEELER, D. K.; SETTE, A.; PETERS, B. The immune epitope database (IEDB): 2018 update. **Nucleic Acids Research**, Oxford, v. 47, p. D339-343, 2018. DOI: 10.1093/nar/gky1006. [Epub ahead of print] PubMed PMID: 30357391.

VITA, R.; ZAREBSKI, L.; GREENBAUM, J. A.; EMAMI, H.; HOOF, I.; SALIMI, N.; DAMLE, R.; SETTE, A.; PETERS, B. The Immune Epitope Database 2.0. **Nucleic Acids Research**, Oxford, v. 38, p. D854–D862. 2010. DOI: https://doi.org/10.1093/nar/gkp1004.

WANG, H. W.; LIN, Y. C.; PAI, T. W.; CHANG, H. T. Prediction of B-cell linear epitopes with a combination of support vector machine classification and amino acid propensity identification. **Journal Biomed Biotechnology**, 2011. DOI: https://doi.org/10.1155/2011/432830.

WEE, L. J.; SIMARMATA, D.; KAM, Y. W. *et al.* SVM-based prediction of linear B-cell epitopes using bayes feature extraction. **BMC Genomics**, London, v. 11, p. S21, 2010. DOI: https://doi.org/10.1186/ 1471-2164-11-S4-S21.

WELLING, G. W.; WEIJER, W. J.; VAN DERZEE, R.; WELLING-WESTER, S. **Amino acid Scale: antigenicity value X 10**. ProtScale, 1985. Disponível em: ProtScale Tool: Antigenicity value X 10. (expasy.org). Acesso em: 01 jun. 2021.

YANG, X.; YU, X. An introduction to epitope prediction methods and software. **Reviews in Medical Virology**, Oxford, v. 19, p. 77–96, 2009. DOI: https://doi.org/10.1002/rmv.602.

YAO, B.; ZHANG, L.; LIANG, S.; ZHANG, C. SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity. **PLoS**

One, San Francisco, v. 7, n. 9, e45152, 2012. DOI: https://doi.org/10.1371/journal.pone.0045152.

ZHAO, Bo. **Web scraping**. Encyclopedia of big data, p. 1-3, 2017.

APÊNDICE A – CURVA ROC DA VALIDAÇÃO CRUZADA DE 5 VEZES (POR EXECUÇÃO) SOBRE CADA CONJUNTO DE DADOS DE TREINAMENTO/ VALIDAÇÃO