Uma Ferramenta Baseada em Redes Neurais da Família ART para Predição de Epítopos Lineares de Células B

Anthony F. La Marca[1], Bruno S. de Oliveira[1], Robson da S. Lopes[1]

[1] Campus Araguaia – Universidade Federal do Mato Grosso – (UFMT)
Avenida Valdon Varjão, 6390 - CEP: 78600-000 - Barra do Garças – MT– Brasil

{anthony.marca,robson.lopes}@ufmt.br, brunoliveira.principal@gmail.com

Abstract. *The public health system relies on the use of vaccines to immunize the population against a range of infectious diseases. To develop and monitor these diseases, precise diagnostic methods are employed, which identify immunogenic regions within a protein. This process is extremely demanding and crucial, making it timely to develop tools that can assist. The present proposal uses the ARTMAP-FUZZY Artificial Neural Network (ANN), trained with epitope-annotated amino acid sequences of proteins available in the IEDB. The data were pre-processed using the amino acid propensity scale and its proportion in positive and negative epitopes. For training and testing, five-fold cross-validation and the area under the ROC curve (AUC) were used to evaluate the results, achieving a value of 0.9289.*

Resumo. *O sistema público de saúde é dependente do uso de vacinas para imunizar a população de uma série de doenças infecciosas. Para desenvolvê-las e monitorar essas doenças utiliza-se métodos de diagnóstico precisos, que identificam regiões imunogênicas dentro de uma proteína. Esse processo é extremamente exigente e crucial, o que torna oportuno desenvolver ferramentas que possam o auxiliar. A presente proposta utiliza a Rede Neural Artificial (RNA) ARTMAP-FUZZY, treinada com epítopos anotados de sequências de aminoácidos de proteína, disponíveis no IEDB. Os dados foram pré-processados utilizando a escala de propensão de aminoácidos e sua proporção em epítopos positivos e negativos. Para o treinamento e teste, foi utilizado validação cruzada quíntupla e a área sob a curva (AUC) ROC para avaliar os resultados, atingindo um valor de 0,9289.*

1. Introdução

Para evitar o colapso do sistema de saúde e mortes em massa, é crucial desenvolver diagnósticos precisos e vacinas eficazes contra doenças infecciosas. Para isso, é preciso identificar as regiões imunogênicas, que representam a interface entre a infecção e a resposta imune, mais conhecido como epítopos de células B [Kringelum 2013].

Os epítopos de células B são classificados como lineares (aminoácidos contínuos) ou conformacionais (aminoácidos contínuos devido ao dobramento da proteína) [Van Regenmortel 2009]. Este estudo concentra-se na previsão de epítopos lineares de células B.

Métodos *in sílico* têm sido amplamente adotados para a previsão de epítopos, principalmente em função dos altos custos associados aos métodos laboratoriais tradicionais. A simplicidade na geração de embeddings tem facilitado o avanço nesse campo nos últimos anos, com ferramentas como BepiPred-2.0 [Jespersen 2017] e EpiDope [Collatz 2020] se destacando. No entanto, o desempenho dessas ferramentas ainda é limitado, o que compromete a precisão na identificação de epítopos de células B [Sun 2019].

Neste contexto, o objetivo é apresentar uma nova ferramenta computacional capaz de identificar e localizar epítopos lineares de células B em sequências de aminoácidos de proteínas. A base de dados foi extraída do IEDB, incluindo a redução de similaridade [Camacho 2013], 14 escalas de propensão de aminoácidos [Lin 2013], a proporção de resíduos em epítopos positivos e negativos, além da técnica de validação cruzada. Os dados foram analisados utilizando uma janela deslizante de tamanho pré-definido, e os resultados foram processados pela RNA ARTMAP-FUZZY. Para a avaliação dos testes, foram utilizadas as métricas de Sensibilidade, Especificidade, Precisão, Acurácia, Coeficiente de Correlação de Matthews (MCC) e Área sob a Curva (AUC) ROC.

2. Materiais e Métodos

2.1 Conjunto de dados de Treinamento e Teste

A primeira etapa desta pesquisa envolveu a busca por epítopos lineares de células B disponíveis em bancos de dados públicos. Dada a sua quantidade, diversidade, validação experimental e os diversos filtros de pesquisa refinados, o banco de dados do IEDB foi selecionado como a fonte para fornecer tanto os epítopos positivos quanto os negativos para os experimentos de treinamento e teste.

A obtenção dos epítopos do IEDB foi realizada em 5 de abril de 2023, resultando em 11.509 epítopos positivos e 28.080 epítopos negativos, todos lineares, referentes a doenças infecciosas em humanos e abrangendo todos os táxons, conforme apresentado na Tabela 1.

Tabela 1 – Quantidade de Epítopos Positivos/Negativos extraídos do IEDB

Táxon	Epítopos Positivos	Epítopos Negativos	Epítopos Positivos Pré Processados	Epítopos Negativos Pré Processados
Bactéria	1803 (15,67%)	4167 (14,84%)	1600 (16.05%)	2300 (21.36%)
Vírus	5569 (48,39%)	6638 (23,64%)	4376 (43.90%)	4475 (41.57%)
Protozoário	4137 (35,94%)	17275 (61,52%)	3992 (40.05%)	3991 (37.07%)
Total	11509	28080	9968	10766

Foram geradas quatro bases de dados: três independentes, cada uma composta por resíduos de um táxon específico, e uma dependente, que é a união dos resíduos de todos os três táxons. O número de epítopos positivos e negativos em cada base é apresentada na Tabela 1 e são nomeadas da seguinte forma: DB_bac, que contém resíduos de epítopos positivos e negativos de bactérias; DB_vir, composta por resíduos de epítopos positivos e negativos de vírus; DB_prot, que inclui resíduos de epítopos positivos e negativos de protozoários; e DB_all, que abrange resíduos de epítopos positivos e negativos de todos os três táxons.

2.2 Preparação dos Dados

Os epítopos com similaridade igual ou superior a 80%, avaliados pelo software BLAST [Camacho 2013], foram agrupados, e uma sequência foi selecionada aleatoriamente de cada grupo para compor o conjunto de dados final. Esse procedimento é essencial para evitar que o algoritmo de aprendizado de máquina memorize sequências de epítopos muito semelhantes, favorecendo assim uma melhor generalização.

Foi observado um grande desbalanceamento entre as classes. Para evitar que os modelos gerados apresentassem viés e que epítopos importantes fossem descartados, foram

aplicadas técnicas de correção de prevalência de forma ponderada, como a Amostragem Estratificada. Além disso, identificou-se que alguns registros continham dados importantes faltantes, os quais foram eliminados. Também foram removidos os registros com epítopos com comprimento superior a 30 ou inferior a 5 aminoácidos, pois esses eram esporádicos na base de dados. Ao final, a base de dados resultante ficou composta por 9.968 epítopos positivos e 10.766 epítopos negativos, conforme detalhado por táxon na Tabela 1.

Esse processo gerou dados sobre a posição inicial e final de cada epítopo, bem como a sequência do antígeno. Além disso, dados adicionais sobre as proteínas foram obtidos da base de dados NCBI (2023), a fim de auxiliar os algoritmos na identificação de correlações.

2.3 Estratégia de Predição

A estratégia de predição adotada foi inspirada na ferramenta BEEPro [Lin 2013], que apresentou bons resultados. Utilizou-se a escala de proporção de aminoácidos, além de 14 propriedades físico-químicas e bioquímicas, obtidas do banco de dados AAindex [Lin 2013].

A taxa de proporção de cada um dos vinte aminoácidos foi calculada pela equação $P_{\alpha i}\frac{f_{\alpha i}^{+}/\sum i f_{\alpha i}^{+}}{f_{\alpha i}^{-}/\sum i f_{\alpha i}^{-}}$, onde $f_{\alpha i}^{+}$ representa a frequência do aminoácido αi em epítopos positivos e $f_{\alpha i}^{-}$ em epítopos negativos. Todos os valores foram então normalizados entre [0, 1] para evitar viés nos algoritmos.

Com esses dados e os valores das 14 propriedades físico-químicas e bioquímicas selecionadas, de acordo com Lin (2013), os epítopos de interesse foram percorridos, gerando os atributos de entrada para o algoritmo de aprendizado de máquina. Para isso, foi usado o método da janela deslizante, onde uma janela móvel atribui uma média ao aminoácido central da janela para cada propriedade j (j = 0, 1, ..., 14), calculada pela equação: $mediaEscala_j = \frac{\sum_i (1 - f * |c - i|) * S_i}{w}$, onde 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) e S_i: valor da propriedade físico-química ou taxa de proporção de aminoácido do resíduo na posição i.

O tamanho da janela deslizante foi determinado pela média do comprimento dos epítopos e por testes empíricos feitos com tamanhos de 10, 12, 15, 17 e 20. A janela de tamanho 20 apresentou os melhores resultados.

2.4 Proposta da RNA ARTMAP-FUZZY

Embora a RNA ARTMAP-FUZZY ofereça boas taxas de generalização, seu desempenho depende das escolhas adequadas dos parâmetros de vigilância: $\rho_a, \rho_b \; e \; \rho_{ab}$. Uma escolha inadequada pode comprometer a acurácia dos resultados; valores próximos de zero permitem que padrões pouco semelhantes sejam agrupados na mesma categoria, enquanto valores próximos de um fazem com que pequenas variações nos padrões de entrada levem a RNA a criar novas classes [Grossberg 2013]. Para este estudo, foram utilizados os valores de 0,61 (ρabaseline), 0.8 e 0.99 para os parâmetros $\rho_a, \rho_b \; e \; \rho_{ab}$, respectivamente. Os parâmetros $\alpha \; e \; \beta$ foram fixados em 0.1 e 1.0, respectivamente, enquanto que a taxa de incremento do ρ_{ab} foi definida empiricamente e fixada em 0.1.

As matrizes pesos $\boldsymbol{w_a}, \boldsymbol{w_b} \; e \; \boldsymbol{w_{ab}}$ foram iniciadas com o valor igual a 1, indicando que todas as atividades estam inativas. Essas matrizes começaram apenas com 1 linha, indicando a existência de apenas um neurônio ativo no início do treinamento. À medida que o processo de aprendizado prosseguia e as atividades eram ativadas, novos neurônios foram criados e inicializados dinamicamente.

2.5 Treinamento

A ordem de apresentação dos padrões de entrada e saída à RNA pode ser feita de forma sequencial ou aleatória/pseudoaleatória. No entanto, sob a ótica cognitiva, é mais eficiente adotar a ordem aleatória/pseudoaleatória, razão pela qual essa abordagem foi escolhida.

Com o uso 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* gerado pela validação cruzada, são gerados três modelos de conhecimento (técnica de competição). Em cada treinamento, os padrões de entrada e saída foram apresentados de maneira pseudoaleatória e em ordens diferentes, permitindo que cada modelo aprendesse de forma distinta, gerando diferentes quantidades de neurônios.

A Tabela 2 apresenta as quantidades de neurônios criados em cada modelo de conhecimento, para cada partição de dados, após o processo de treinamento e validação cruzada. Vale destacar, que essas quantidades se referem apenas as matrizes de peso w_a e w_{ab}, visto que a matriz peso w_b possui apenas dois neurônios.

Tabela 2 - 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 Neurônios DB_bac	Quantidade Neurônios DB_vir	Quantidade Neurônios DB_prot	Quantidade 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

2.5 Diagnóstico

No diagnóstico, a RNA ativa a categoria que melhor representa o padrão de entrada por meio da função de escolha, verificando se o grau de similaridade entre o padrão de entrada e a categoria ativada atende ao valor definido pelo parâmetro ρ_{a}. No entanto, pode haver padrões

de entrada com características muito distintas das categorias já existentes, de modo que nenhuma delas seja suficientemente similar para passar no teste de vigilância.

Para o processo de diagnóstico foi utilizado validação cruzada e a estratégia de competição, na qual cada modelo de conhecimento gerado (três) 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.

3. Resultados

Após o processo de diagnóstico e de posse dos resultados das bases de dados DB_bac, DB_vir, DB_prot e DB_all, foram aplicadas as métricas: sensibilidade, especificidade, precisão, acurácia e MCC. Os resultados dessas métricas estão apresentadas na Tabela 3.

Tabela 3 - 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

Para avaliar a capacidade do método em identificar epítopos positivos, foi calculada a média da curva ROC para cada subconjunto de cada base de dados. Utilizando validação cruzada quíntupla, uma curva ROC foi gerada para cada partição, e, ao final, a média dessas curvas foi calculada. A Figura 1 apresenta as médias das curvas ROC para as bases de dados DB_bac (a), DB_vir (b), DB_prot (c) e DB_all (d).

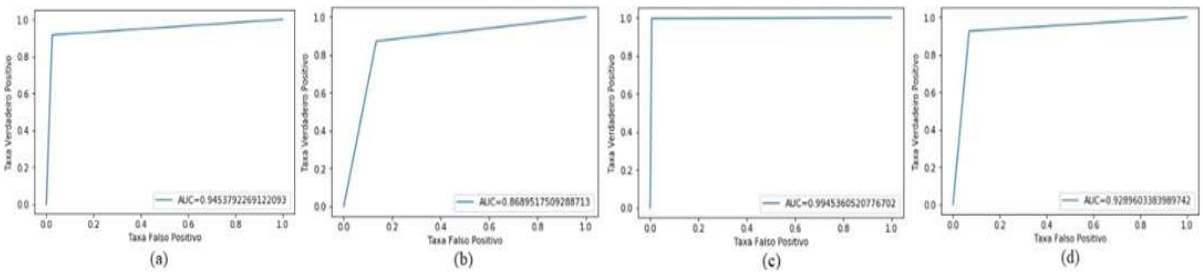

Figura 1 - Área sob a curva ROC utilizando validação cruzada de 5 vezes para os Conjuntos de Dados: DB_bac (a), DB_vir (b), DB_prot (c) e DB_all (d)

Todos os conjuntos usados apresentaram uma curva ROC com acurácia superior a 86%. A base de dados de protozoários (DB_prot) obteve o melhor desempenho, com uma acurácia de aproximadamente 99,45%. O menor índice foi observado na base viral (DB_vir), com cerca de 86,89%, refletindo sua alta taxa de mutação. As demais bases ficaram dentro dessa faixa, com acurácias de aproximadamente 94,54% para DB_bac e 92,89% para DB_all.

4. Considerações Finais

Este estudo apresentou uma ferramenta para predição de epítopos lineares de células B, que, apesar de requerer apenas a sequência de aminoácidos de uma proteína como entrada, demonstrou um desempenho aceitável no processo de predição.

Todas as proteínas utilizadas nos conjuntos de dados de treinamento e teste possuem similaridade inferior a 80%, o que assegura a independência entre os conjuntos e aumenta a capacidade de generalização da RNA ARTMAP-FUZZY.

O desempenho da ferramenta pode ser creditado à estabilidade e plasticidade inerentes às redes da família ART, à sua forte capacidade de generalização para esses dados, e à técnica de competição aplicada nos testes. Além disso, com um pré-processamento de baixo custo computacional, há viabilidade para o desenvolvimento de uma versão Web.

Um dos principais desafios enfrentados foi definir com precisão os valores ideais para os parâmetros de entrada, tanto da RNA quanto os utilizados no pré-processamento. Pequenas variações nesses parâmetros resultavam em diferenças significativas nos resultados.

Para trabalhos futuros, destaca-se o desenvolvimento de uma interface gráfica intuitiva que permita aos profissionais da área utilizá-la amplamente em suas atividades diárias. Além disso, ressalta-se a importância de atualizações constantes na base de dados de treinamento, visando aprimorar continuamente os *insights* da RNA e sua otimização.

As ferramentas de predição de epítopos devem atuar principalmente como filtros para descartar regiões improváveis de serem epítopos e, assim, eliminar análises experimentais desnecessários. Para isso, é essencial que essas ferramentas ofereçam sensibilidade e especificidade adequadas, tornando os experimentos mais precisos e direcionados.

Referências

Kringelum, J. V., Nielsen, M., Padkjær, S. B., Lund, O. (2013). Structural analysis of b-cell epitopes in antibody: protein complexes. https://doi.org/10.1016/j.molimm.2012.06.001.

Van Regenmortel, M. H. (2009) What is a b-cell epitope? In Epitope Mapping Protocols, pages 3-20. https://doi.org/10.1007/978-1-59745-450-6_1.

Sun, P., Guo, S., Sun, J., Tan, L., Lu, C., Ma, Z. (2019). Advances in In-silico B-cell Epitope Prediction. https://doi.org/10.2174/1568026619666181130111827.

Jespersen, M. C., Peters, B., Nielsen, M., Marcatili, P. (2017). BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes**.** Nucleic acids research. https://doi.org/10.1093/nar/gkx346.

Collatz, M., Mock, F., Hölzer, M., Barth., E., Sachse, K., Marz, M. (2020). EpiDope: A Deep neural network for linear B-cell epitope prediction. https://doi. org/10.1101/2020.05.12.090019.

Camacho, C., Madden, T., Coulouris, G., Avagyan, V., Ma, N., Tao, T., Agarwala, R. (2013). BLAST Command Line Applications User Manual. http://nebc.nerc.ac.uk/bioinformatics/ documentation/blast+/user_manual.pdf.

Lin, S. H., Cheng, C. W., Su, E. C. (2013). Prediction of B-cell epitopes using evolutionary information and propensity scales. https://doi.org/10.1186/1471-2105-14-S2-S10.

NCBI – National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov.

Grossberg, S. Adaptive resonance theory: how a brain learns to consciously attend, learn, and recognize a changing world. Neural Networks. 2013. Doi: 10.1016/j.neunet.2012.09.017