Investigando a relação entre os aminoácidos de proteínas do vírus da dengue e o desfecho clínico do paciente

Diego Queiroz[1], Fagner Cunha[1], Leonardo Rodrigues Souza[1], Juan G. Colonna[1]

[1]*Instituto de Computação, Universidade Federal do Amazonas, Manaus, AM*

{diego.queiroz, fagner.cunha, rdsouza.leonardo, juancolonna}@icomp.ufam.edu.br

Abstract. *In this work we propose a simplified method to represent dengue virus proteins and classify them according to the severity of the infection, which are classic and severe. These classes identify the clinical outcome of the patient, allowing to link the genomic composition of the virus and the reaction it caused in patients. To accomplish this, we transformed the protein sequences into a set of complex networks (graphs), from which histograms with the degree of nodes were generated. The representations were classified by a Decision Tree. For validation, the Leave-One-Out method was used. The classifier reached an AUC between 70% to 84%*

Resumo. *Neste trabalho propomos um método simplificado para representar proteínas do vírus da dengue e classificá-las de acordo com a severidade da infecção, sendo estas clássica e severa. Essas classes identificam o desfecho clínico do paciente, permitindo relacionar a composição genômica do vírus e a reação que esse causou nos pacientes. Para isso, transformamos as sequências proteicas em um conjunto de redes complexas (grafos), a partir das quais foram gerados histogramas com o grau dos nós. As representações foram classificadas por uma Árvore de Decisão. Para validação empregou-se o método Leave-One-Out. O classificador atingiu uma área média sob a curva ROC de 70% a 84%.*

Palavras-chave: Bioinformática; Vírus da dengue; Redes Complexas.

1. Introdução

Sequências genômicas completas (RNA/DNA), assim como as proteínas que se traduzem a partir dessas, são úteis para identificar tratamentos, vacinas e drogas para diversas doenças. Sendo assim, podemos hipotetizar que as sequências proteicas podem possuir padrões genômicos que estejam associados com o estado clínico do paciente.

O objetivo deste trabalho é encontrar, a partir da simplificação da metodologia desenvolvida por Ito et al (2017), padrões presentes nas sequências proteicas do vírus da dengue que ajudem a identificar os aminoácidos mais significativos relacionados ao desfecho clínico do paciente com dengue. Para isso, as sequências proteicas foram mapeadas em grafos (ou redes complexas) que permitem estabelecer relações entre os

aminoácidos. Após isso, foram extraídas características dos grafos que permitiram classificar essas sequências em dengue severa ou clássica.

Nossa contribuição é uma metodologia simplificada baseada unicamente na utilização do grau dos nós do grafo (*kernel degree*), em contraste com a metodologia original que utilizou 10 atributos extraídos dos grafos. Além de reduzir a complexidade e melhorar os resultados, a nova metodologia também permite identificar para cada sequência os aminoácidos relevantes para classificar a dengue severa. Os novos resultados confirmam a hipótese de que a relação entre aminoácidos vizinhos nas sequências carrega informações relevantes que caracterizam a doença.

2. Trabalhos Relacionados

Alguns trabalhos têm aplicado técnicas de aprendizado de máquina para estudar a classificação da dengue. Iqbal e Islam (2019) compararam vários modelos para classificar surtos de dengue utilizando oito atributos: febre, enxaqueca, dor corporal, dor abdominal, vômito, hemoglobina, WBC e plaquetas, com os modelos *LogitBoost* e o *Random Forest* apresentando os melhores desempenhos com 92% de acurácia. Balasaravanan e Prakash (2018) utilizaram uma rede neural artificial com quatro atributos como entrada (chuva, taxa de umidade, temperatura e o número de casos de dengue do conjunto de dados do mês anterior) para detectar ocorrência de dengue no paciente, atingindo uma acurácia de 92%.

O trabalho tomado como *baseline* para esta pesquisa foi o de Ito et al (2017), onde foi realizado um estudo a partir de sequências de RNA de variadas espécies de animais e plantas com o objetivo de classificá-las em RNA não-codificante e RNA mensageiro. Primeiramente, para cada sequência, representou-se a relação de proximidade entre os códons, transformando-as em grafos. Após isso, foi extraído um conjunto de 10 atributos: a assortatividade, os grau médio, máximo e mínimo de um vértice, a média do caminho mínimo, o coeficiente de agrupamento, a centralidade da intermediação média, o desvio padrão, e as frequência de subgrafos de tamanho 3 e de tamanho 4. O modelo aplicado foi uma árvore de decisão e atingiu uma acurácia média de 99,79%.

No entanto, em nossa adaptação, utilizamos o *Degree Kernel* para simplificar a metodologia proposta por Ito et al (2017). Nossa metodologia, por fazer uso de histogramas com os graus de cada nó normalizados, ajuda na identificação dos nós que possuem mais ligações na rede complexa, e assim, identificar os aminoácidos mais significantes para classificação do quadro clínico que o paciente desenvolveu.

3. Fundamentos Teóricos

Um aminoácido é codificado por uma sub-sequência de três nucleotídeos (códons) que, por sua vez, geram as sequências proteicas. Os códons do RNA são combinações dos nucleotídeos adenina (A), citosina (C), guanina (G) e uracila (U) [Souza 2022].

Uma rede complexa (RC) é um grafo no qual as arestas podem ter pesos. As arestas estabelecem algum tipo de relação entre dois nós de acordo com o problema modelado [Jean Metz et al. 2007]. Na teoria dos grafos, o grau de um vértice é o número de arestas incidentes nele, com os laços contados duas vezes [Diestel 2005]. Portanto, o

grau do nó ajuda a identificar qual códon da sequência tem mais influência. Dado um grafo G = (V, E), o grau de um nó *v* é dado por:

$$\sum deg(v) = 2|E|.$$

A média do caminho mínimo é um conceito na topologia das redes que é definido como o número médio de passos que um nó precisa percorrer para chegar em seu par pelo caminho mais curto [Guoyong Mao and Ning Zhang 2013]. O coeficiente de agrupamento mede o grau com que os nós de um grafo tendem a agrupar-se [P. W. Holland and S. Leinhardt 1971]. No grafo de uma sequência proteica, os grupos são compostos por aminoácidos que possuem características parecidas. A centralidade quantifica o número de vezes que um nó age como ponte ao longo do caminho mais curto [Freeman 1977]. Esse atributo identifica o aminoácido que serve de nó comunicador entre aqueles que não estão próximos, para que haja interação da sequência por completo. A assortatividade quantifica a tendência de nós individuais se conectarem a outros nós semelhantes em um grafo [Foster 2011]. Nas estruturas proteicas, quantifica a tendência de conexão entre aminoácidos com atributos semelhantes. O desvio padrão mede o grau de variância de um conjunto de elementos [Bland 1996]. Em grafos, representa coesão dos nós.

4. Materiais e Métodos

4.1. Base de Dados

A base de dados utilizada neste trabalho foi disponibilizada por Souza (2022), que foi escolhida por estar previamente tratada e separada por proteínas, o que facilitou no andamento do estudo. A **Tabela 1** apresenta a distribuição de amostras em cada proteína de acordo com rótulos: dengue severa e dengue clássica.

Tabela 1. Quantidade de amostras por rótulo de cada proteína.

Proteína	C	E	M	NS1	NS2A	NS2B	NS3	NS4A	NS4B	NS5
Dengue Clássica	71	190	91	110	104	70	132	85	89	148
Dengue Severa	35	81	42	47	34	33	50	40	37	62
Total	106	271	133	157	138	103	182	125	126	210

4.2. Geração de Rede Complexa

A **Figura 1** ilustra as duas metodologias (atributos da rede complexa e *Degree Kernel*) e onde elas se divergem, identificando os três passos principais do processo: mapeamento, extração de características e classificação. O mapeamento consiste em mapear as sequências proteicas no formato FASTA e em grafos para, por fim, realizar a extração das características. Os grafos das sequências foram gerados de acordo com o mapeamento original proposto por Ito et al (2017), que possuem WS = 3 (*Word Size*) e SS = 1 (*Step Size*). O WS corresponde ao tamanho dos aminoácidos que geram os nós da rede e o SS é a distância entre pares de vizinhos. Com essas informações, é possível identificar os nós que se repetem e qual a relação entre eles. Assim, é possível atribuir pesos às arestas para reconhecer a importância de cada ligação. No entanto, podem haver erros de transcrição do RNA, então, para todo caractere diferente de A, C, G ou U

ocorre a substituição pela incógnita X. Com isso, a quantidade possível de códon (nós no grafo) é reduzida para 125. Vale ressaltar que nossa implementação da extração de atributos da rede complexa não inclui os atributos de frequência de subgrafos de tamanho 3 e de tamanho 4, pois não havia implementação funcional em python.

O *Degree Kernel* gera um histograma com o grau dos nós (número de arestas incidentes a um determinado nó). Em seguida, o conjunto de histogramas é utilizado para treinar uma Árvore de Decisão. Os classificadores baseados em Árvores de Decisão possuem uma forma de atribuir importância às características dos vetores de treinamento, dessa forma, podemos identificar os aminoácidos mais importantes de cada proteína e relacionar o desfecho clínico de acordo com os aminoácidos mais importantes do grafo.

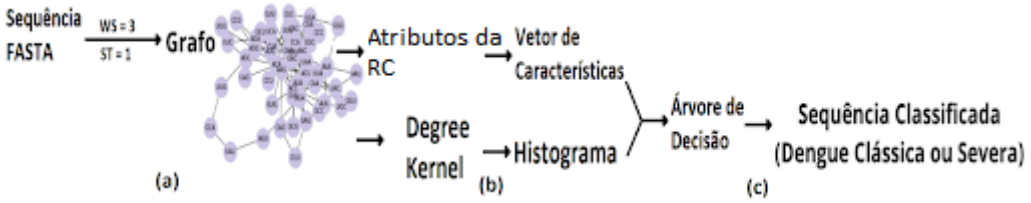

Figura 1. Metodologia a qual foi separada em mapeamento (a), extração de características (b) e classificação (c).

4.3. Classificação

O classificador utilizado foi a Árvore de Decisão, que também é empregada no trabalho de Ito et al (2017). Este modelo foi escolhido por ser simples e ter a vantagem de produzir um modelo de classificação compreensível com níveis de precisão satisfatórios [Sani H.M. 2018]. O parâmetro de profundidade da árvore escolhida foi de 3. Esse parâmetro foi obtido após otimização obtida via *Grid Search Cross Validation*. O método de validação utilizado foi o *Leave-One-Out.*

5. Resultados

A **Figura 2** mostra as curvas ROC da proteína NS2B da nossa metodologia e da metodologia de Ito et al (2017). As curvas mostram a relação entre a taxa de falsos positivos e a taxa de verdadeiros positivos, portanto, quanto maior é a área sob a curva, melhor é o desempenho do classificador. Nessa figura podemos observar que a nossa metodologia obteve um desempenho melhor.

Como resultado da simplificação proposta, separamos duas proteínas, uma que contém o maior número de amostras, proteína E, e outra que contém o menor, proteína NS2B, para calcular o tempo de execução. A proteína E teve um tempo de execução de aproximadamente 166 segundos, e a proteína NS2B, 11 segundos.

Conforme mostra a **Tabela 2**, podemos observar que na maioria das proteínas as metodologias obtiveram resultados aproximados, com divergências em algumas proteínas específicas, mas de forma geral a metodologia com degree kernel foi melhor**.** Utilizando a ferramenta *feature importance* conseguimos verificar que o aminoácido GAU (Ácido aspártico) se destaca na classificação da severidade da dengue para amostras da proteína NS2B.

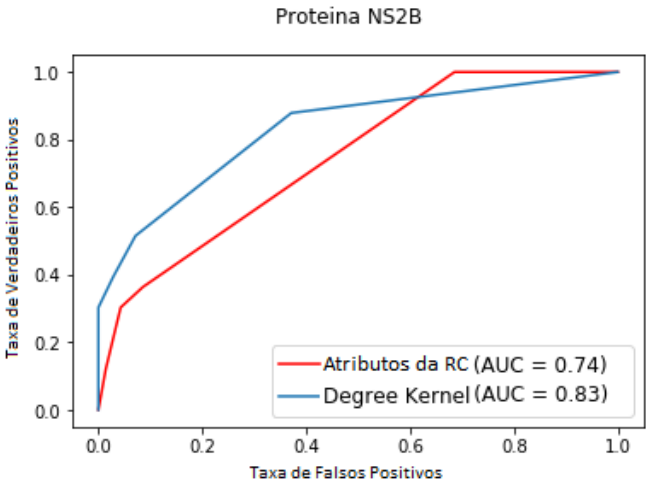

Figura 2. Gráfico mostrando curvas ROC e AUC da proteína NS2B das duas metodologias.

Tabela 2. Resultados obtidos através da técnica de Leave-One-Out.

	Atributos da Rede Complexa			Degree Kernel		
	Acurácia	**Curva ROC**	**Curva PR**	**Acurácia**	**Curva ROC**	**Curva PR**
Proteína E	75,27%	0.76	0.57	**76,01%**	**0.79**	**0.66**
Proteína C	65,09%	**0.87**	0.74	**66,03%**	0.82	**0.75**
Proteína M	57,89%	0.80	0.66	**66,16%**	**0.84**	**0.68**
Proteína NS1	57,96%	**0.81**	0.62	**76,43%**	0.77	**0.69**
Proteína NS2A	76,08%	**0.84**	0.62	**78,26%**	0.81	**0.63**
Proteína NS2B	72,81%	0.74	0.54	**74,75%**	**0.83**	**0.70**
Proteína NS3	70,87%	**0.81**	**0.57**	**71,42%**	0.73	0.49
Proteína NS4A	**69,60%**	**0.83**	**0.61**	56,00%	0.80	0.59
Proteína NS4B	63,49%	**0.83**	**0.65**	**67,46%**	0.70	0.54
Proteína NS5	**63,80%**	0.74	0.53	60,95%	**0.76**	**0.54**

6. Conclusões

Por meio do método proposto e utilizando a ferramenta *feature importance,* mostramos que é possível identificar o aminoácido mais relevante de cada proteína. Com esses resultados, concluímos que, apesar de menos complexa, a metodologia proposta obteve resultados aproximados aos da metodologia de Ito et al (2017) para classificação de proteínas da dengue de acordo com o desfecho clínico de pacientes. Além disso, nossa metodologia também proporciona a interpretação do classificador ao utilizar o histograma com graus dos nós normalizados. Para trabalhos futuros, pretendemos verificar quais aminoácidos estão ligados com o aminoácido mais importante, a fim de saber como ele influencia nas ligações presentes em cada proteína. Também desejamos empregar um interpretador local de modelos com o objetivo de identificar aminoácidos unicamente relevantes para a classificação de dengue severa.

Agradecimentos

Este artigo foi produzido no âmbito do Projeto Samsung-UFAM de Ensino e Pesquisa (SUPER), conforme previsto no Artigo 48 do Decreto nº 6.008 / 2006 (SUFRAMA), que foi financiada pela Samsung Eletrônica da Amazônia Ltda., nos termos da Lei Federal nº 8.387 / 1991, através do convênio 001/2020, firmado com a Universidade Federal do Amazonas e a FAEPI, Brasil.

Referências

Balasaravanan K. and Prakash M. (2018). “Detection of dengue disease using artificial neural network based classification technique”. International Journal of Engineering & Technology, 7 (1.3) (2018) 13-15.

Bland, J.M., Altman, D.G. (1996). “Statistics notes: measurement error”. BMJ. 312 (7047): 1654. doi:10.1136/bmj.312.7047.1654. PMC 2351401. PMID 8664723.

Diestel, R. (2005). “Graph Theory” 3ª ed. Berlin, New York: Springer-Verlag. ISBN 978-3-540-26183-4.

Foster, D.V., Foster, J.G.,Grassberger, P., Paczuski, M. (2011). “Clustering drives assortativity and community structure in ensembles of networks”. Physical review. E, Statistical, nonlinear, and soft matter physics. 84 (6 Pt 2). 066117 páginas. PMID 22304165. doi:10.1103/PhysRevE.84.066117.

Freeman, L. (1977). "A set of measures of centrality based on betweenness". Sociometry. 40(1): 35–41. doi:10.2307/3033543. JSTOR 3033543.

Holland, P.W. and Leinhardt, S. (1971). "Transitivity in structural models of small groups". Comparative Group Studies 2: 107–124.

Ito, E.A., Katahira, I., Vicente, F.F.R., Pereira, L.F.P., and Lopes, F.M. (2017). “BASiNET—BiologicAl Sequences NETwork: a case study on coding and non-coding RNAs identification”. Nucleic Acids Research, 2018, Vol. 46, No. 16.

Iqbal N. and Islam M. (2019). “Machine Learning for Dengue Outbreak Prediction: A Performance Evaluation of Different Prominent Classifiers”. Informatica 43 (2019) 363–371. https://doi.org/10.31449/inf.v43i1.1548

Mao, G. and Zhang, N. (2013) “Analysis of Average Shortest-Path Length of Scale-Free Network”. doi: 10.1155/2013/865643.

Metz, J., Calvo, R., Seno, E.R.M., Romero, R.A.F., Liang, Z. (2007). “Redes Complexas: conceitos e aplicações”. Instituto de Ciências Matemáticas e de Computação, No 290, São Carlos, 2007, páginas 9--21, issn: 0103-2569

Sani, H.M., Lei, C., Neagu, D. (2018). “Computational Complexity Analysis of Decision Tree Algorithms”. In: Bramer, M., Petridis, M. (eds) Artificial Intelligence XXXV. SGAI 2018. Lecture Notes in Computer Science, vol 11311. Springer, Cham.

Souza, L.R., Colonna, J.G., Comodaro, J.M. (2022) “Using amino acids co-occurrence matrices and explainability model to investigate patterns in dengue virus proteins”. BMC Bioinformatics 23, 80. https://doi.org/10.1186/s12859-022-04597-y