Nova Proposta de Representação de Genoma Viral Aplicada na Classificação do SARS-CoV-2 com Aprendizagem Profunda

Luísa C. de Souza* , Marcelo A. C. Fernandes*†‡
*Laboratório de Aprendizagem de Máquina e Instrumentação Inteligente (LAMII), nPITI/IMD, UFRN, Natal, RN, Brasil.
*Centro Multiusuário de Bioinformática (BioME) - IMD, UFRN, Natal, RN, Basil.
‡Departamento de Engenharia da Computação e Automação (DCA), UFRN, Natal, RN, Basil.
Email: luisa.souza.103@ufrn.edu.br, mfernandes@dca.ufrn.br

Informações Gerais

Título da Dissertação: Nova Proposta de Representação de Genoma Viral Aplicada na Classificação do SARS-CoV-2 com Aprendizagem Profunda.
Candidata ao Mestrado: Luísa Christina de Souza.
Orientador: Professor Dr. Marcelo Augusto Costa Fernandes.
Departamento: Programa de Pós-Graduação em Engenharia Elétrica e de Computação.
Instituição: Universidade Federal do Rio Grande do Norte (UFRN).

I. Motivação

A Organização Mundial de Saúde (OMS) declarou em 30 de janeiro de 2020, que o surto do COVID-19 constituía uma Emergência de Saúde Pública de Interesse Internacional. A doença causada pelo vírus *Síndrome Respiratória Aguda Grave Coronavírus 2* (SARS-CoV-2) apresentava rápida propagação, tal que após duas semanas do primeiro caso diagnosticado, outros 1000 pacientes testaram positivo para COVID-19 [1]–[3]. O vírus do COVID-19 codifica em seu genoma proteínas estruturais e não estruturais, e entre as estruturais está a proteína Spike, que é o mecanismo utilizado pelo vírus para reconhecer e se acoplar ao receptor *enzima conversora da angiotensina* 2 (ECA2) da célula hospedeira e realizar o processo de fusão com a membrana celular [4], [5]. Em virtude de suas propriedades e funções na infecção viral, esta proteína se tornou um alvo para desenvolvimento de vacinas e terapias [6]. Porém, devido a taxa de mutação que os vírus de RNA apresentam, mudanças nos aminoácidos da proteína spike ou no genoma do vírus, geram variantes do SARS-CoV-2 [4], [7]. Estas podem se manifestar mais resistentes à vacinas e mais transmissíveis [8], [9].

Neste contexto, a classificação, descrição e comparação de sequências virais baseada nas suas características genômicas, podem auxiliar no estudo das relações filogenéticas e mecanismos de atuação dos patógenos, contribuindo para o desenvolvimento de vacinas e de outras medidas de profilaxia [10].As Sequências de DNA complementar (DNAc), contém informações genéticas em suas moléculas que sistematizam o desenvolvimento e funcionamento de organismos vivos e vírus. As sequências formadas por bases de nucleotídeos se apresentam na forma de vetores de caracteres, onde cada letra representa uma base nitrogenada específica, guanina (G), adenina (A), timina (T), e citosina (C) [11].

Na bioinformática a análise desses dados genômicos é realizada através de dois métodos principais. O primeiro método trata-se de técnicas que utilizam alinhamento de sequências, como o BLAST [12] e o BLAT [13]. Tais algoritmos procuram por correspondentes de bases ou grupos de bases na mesma ordem em duas ou mais sequências. As desvantagens apresentadas por essas técnicas são o alto custo de memória e de tempo requeridos, o que limita seu uso em grandes base de dados genômicos [14], além de assumir que as sequências de DNAc são linearmente arranjadas, o que não é o caso para sequências virais. Ademais, a aplicação de tais métodos não se mostra adequada em cenários onde as sequências apresentam grandes divergências ou na comparação de sequências com milhões de nucleotídeos [15]–[17].

O segundo método engloba as técnicas nas quais não é realizado o alinhamento de sequências (*free-alignment*) [18]. Tal método foi desenvolvido como alternativa para solução de problemas biológicos onde as técnicas de alinhamento apresentam limitações. Sendo aplicado em diversos estudos como na análise da evolução de organismos e de sequências de regulação como promotores e inibidores, na identificação de módulos cis-reguladores (CRM) e na comparação de sequências utilizando dados de tecnologias de *next-generation sequencing* [16].

O processamento, análise e transformação de informações genômicas, como as sequências de DNA, quando realizado com técnicas de processamento digital de sinais, é intitulado Processamento de Sinais Genômicos (PSG) [11], [19]. Sendo aplicado na bioinformática predominatemente em dois campos de estudo, no mapeamento do DNA para sinal, com o intuito de transformar os caracteres em informação numérica ou gráfica, e na busca por características e propriedades intrínsecas das sequências de DNA com o uso de ferramentas matemáticas [20].

Nos anos recentes, técnicas de aprendizagem profunda, *Deep learning*, vem sendo amplamente utilizadas na área

de informática da saúde, em análise de imagens médicas, genômica computacional, análise de sinais fisiológicos, representação de dados médicos e predição de doenças, alcançando performance e implementação similar ou superior que técnicas de aprendizagem de máquina tradicionais [21], apresentando inclusive, resultados bastante significativos na área de classificação viral.

Perante isto, trabalho propôs uma nova estratégia de representação de sequências de DNAc viral, tal como a do SARS-CoV-2 e suas novas variantes, utilizando um conjunto de técnicas de processamento de sinais genômicos, tais como a *Chaos Game Representation* (CGR) e a *Discrete Fourier Transform* (DFT), para ser empregada em métodos de aprendizagem profunda para classificação viral. Tal representação de sequências genéticas gera uma nova assinatura viral contendo as informações em um novo espaço de características, e apresentam comprimento consideravelmente menor que a sequência genômica original.

II. Objetivo

O objetivo principal do trabalho foi desenvolver uma nova metodologia de representação de sequências genômicas para um novo espaço de características, apresentando uma assinatura numérica reduzida que ao ser aplicada em técnicas de aprendizagem profunda, foi capaz de diferenciar as espécies virais de modo que o modelo classificou os dados e características genéticas dos vírus, com baixo custo computacional de memória e tempo requerido mantendo altas acurácias.

Além disto, este trabalho também teve como objetivo analisar o comportamento e evolução das variantes virais do SARS-CoV-2 de forma temporal e suas relações entre as linhagens, utilizando a metodologia de representação proposta em conjunto com a arquitetura da rede neural convolucional, e em seguida comparar os resultados obtidos com técnicas de análise da bioinformática convencionais.

III. Contribuições

As abordagens de representação de sequências genéticas e do modelo de classificação viral com aprendizagem profunda desenvolvidas neste estudo, apresentam desempenho similar a outras técnicas de referência, porém com performance superior em relação a custo de memória e de tempo. Baseando-se nas técnicas e resultados dos trabalhos da literatura, as principais contribuições apresentadas pelo método desenvolvido na presente dissertação, são as seguintes:

1) Propomos uma metodologia de representação de sequências virais com ferramentas de PSG, para gerar assinaturas virais reduzidas.
2) Utilizamos a metodologia proposta na classificação do vírus SARS-CoV-2 em um dataset contendo amostras da mesma família do vírus. Sendo útil na discriminação do SARS-CoV-2, que é fortemente relacionado a outras espécies de coronavírus.
3) Na classificação viral, utilizamos uma arquitetura de aprendizagem profunda, e mostramos que, mesmo contendo apenas de 64 a 256 valores no vetor da assinatura viral, o classificador foi capaz de diferenciar entre espécies com alta acurácia.
4) Comparamos o desempenho da representação com técnicas consolidadas na literatura, e mostramos que a abordagem proposta apresenta performance similar ou superior.
5) Aplicamos o método proposto na análise das variantes do SARS-CoV-2.
6) Mostramos, através de análises estatísticas, que os resultados obtidos por técnicas convencionais sustentam os resultados do classificador para a análise das variantes.

IV. Resultados Relevantes

Inicialmente, aplicamos a CGR as sequências genômicas, obtendo coordenadas espaciais que foram aplicadas à DFT, e em comparação com outros trabalhos que utilizam a transformação de Fourier no pré-processamento de amostras de dados genéticos, o presente método utiliza a informação de fase juntamente com a informação de amplitude dos sinais, para aumentar a diferenciação entre amostras.

A redução do tamanho dos vetores de assinatura viral permite uma classificação viral com baixo custo computacional, tanto no tempo de treinamento do modelo de classificação, quanto na quantidade de memória requerida para o armazenamento, características relevantes no tratamento de grandes quantidades de dados, como é o caso das sequências genômicas disponíveis pelas tecnologias de *next-generation sequencing*. Apesar do baixo custo de processamento, o método não teve perda do desempenho, alcançando uma acurácia de $98,9\%$, $97,9\%$ e $99,4\%$, e AUC de $0,9764$, $0,9490$ e $0,9869$ para comprimento do vetor igual à 64, 128 e 256, respectivamente, na classificação realizada com os vírus do SARS-CoV-2 e outras espécies da mesma família, como o *Betacoronarívus 1*, MERS-CoV, HCoV NL63, HCoV 229E e HCoV HKU1.

Após observar o funcionamento da metodologia desenvolvida, esta foi empregada na análise das variantes de preocupação Alpha, Delta, Beta, Gamma e Omicron, considerando o tempo como um fator de mudança do perfil genético das novas estirpes. Os resultados obtidos mostram que, para todas as VOC analisadas individualmente, existe um acréscimo da classificação com a arquitetura de aprendizagem profunda, de forma mais acentuada para as variantes Alpha e Omicron, o que pode indicar um acumulo das mutações nas sequências virais. As classificações em pares, mostram que para todos os tamanhos do vetor de assinatura, a VOC Delta apresenta características intrínsecas suficientes para as diferenciar das demais variantes. Por fim, utilizamos técnicas convencionais de análise de sequências genômicas, para demonstrar que existem entre o conjunto de sequências de DNAc pertencentas a cada variante, diferenças entre os perfis genômicos.

A. Publicações

1) BMC Bioinformatics: A implementação da técnica de representação de dados genômicos proposta neste trabalho foi

publicada após a defesa da dissertação no BMC Bioinformatics Journal que apresenta Fator de Impacto de 3.327 e Qualis A1. Referência do artigo: de Souza, L.C., Azevedo, K.S., de Souza, J.G. et al. New proposal of viral genome representation applied in the classification of SARS-CoV-2 with deep learning. BMC Bioinformatics 24, 92 (2023). https://doi.org/10.1186/s12859-023-05188-1.

2) XV CONGRESSO BRASILEIRO DE INTELIGÊNCIA COMPUTACIONAL: O método proposto no presente trabalho também foi exposto no CBIC 2021 na sessão técnica Inteligência Computacional aplicada ao combate da COVID-19. Referência do artigo: Souza, Luíza De Melo Barbosa, Raquel & Fernandes, Marcelo. (2021). Nova Proposta de Representação de Genoma Viral Aplicada na Classificação do SARS- CoV-2 com Aprendizagem Profunda. 1-8. 10.21528/CBIC2021-76.

B. Defesa da Dissertação

A defesa da dissertação ocorreu no dia 19 de agosto de 2022, com a banca formada pelo examinador interno Professor Dr. Luiz Marcos Garcia Gonçalves (UFRN - Brasil) e pelo examinador externo Dr. Leonardo Alves Dias (University of Birmingham - United Kingdom).

REFERENCES

[1] A. Spinelli and G. Pellino, "Covid-19 pandemic: perspectives on an unfolding crisis," *Journal of British Surgery*, vol. 107, no. 7, pp. 785–787, 2020.

[2] World Health Organization, "Origin of sars-cov-2, 26 march 2020," Technical documents, 2020.

[3] ——, "Coronavirus disease (covid-19)," 2020, https://www.who.int/emergencies/diseases/novel-coronavirus-2019, Last accessed on 2021-01-11.

[4] S. Ali, B. Sahoo, N. Ullah, A. Zelikovskiy, M. Patterson, and I. Khan, "A k-mer based approach for sars-cov-2 variant identification," in *International Symposium on Bioinformatics Research and Applications*. Springer, 2021, pp. 153–164.

[5] K. Kuzmin, A. E. Adeniyi, A. K. DaSouza Jr, D. Lim, H. Nguyen, N. R. Molina, L. Xiong, I. T. Weber, and R. W. Harrison, "Machine learning methods accurately predict host specificity of coronaviruses based on spike sequences alone," *Biochemical and Biophysical Research Communications*, vol. 533, no. 3, pp. 553–558, 2020.

[6] Y. Huang, C. Yang, X.-f. Xu, W. Xu, and S.-w. Liu, "Structural and functional properties of sars-cov-2 spike protein: potential antivirus drug development for covid-19," *Acta Pharmacologica Sinica*, vol. 41, no. 9, pp. 1141–1149, 2020.

[7] S. E. Galloway, P. Paul, D. R. MacCannell, M. A. Johansson, J. T. Brooks, A. MacNeil, R. B. Slayton, S. Tong, B. J. Silk, G. L. Armstrong *et al.*, "Emergence of sars-cov-2 b. 1.1. 7 lineage—united states, december 29, 2020–january 12, 2021," *Morbidity and Mortality Weekly Report*, vol. 70, no. 3, p. 95, 2021.

[8] G. S. Krishnan, S. S. Kamath, and V. Sugumaran, "Predicting vaccine hesitancy and vaccine sentiment using topic modeling and evolutionary optimization," in *International Conference on Applications of Natural Language to Information Systems*. Springer, 2021, pp. 255–263.

[9] M. Ahmad, S. Ali, J. Tariq, I. Khan, M. Shabbir, and A. Zaman, "Combinatorial trace method for network immunization," *Information Sciences*, vol. 519, pp. 215–228, 2020.

[10] M. A. Remita, A. Halioui, A. A. Malick Diouara, B. Daigle, G. Kiani, and A. B. Diallo, "A machine learning approach for viral genome classification," *BMC bioinformatics*, vol. 18, no. 1, pp. 1–11, 2017.

[11] H. K. Kwan and S. B. Arniker, "Numerical representation of dna sequences," in *2009 IEEE International Conference on Electro/Information Technology*. IEEE, 2009, pp. 307–310.

[12] A. Lopez-Rincon, A. Tonda, L. Mendoza-Maldonado, D. G. Mulders, R. Molenkamp, C. A. Perez-Romero, E. Claassen, J. Garssen, and A. D. Kraneveld, "Classification and specific primer design for accurate detection of sars-cov-2 using deep learning," *Scientific reports*, vol. 11, no. 1, pp. 1–11, 2021.

[13] W. J. Kent, "Blat—the blast-like alignment tool," *Genome research*, vol. 12, no. 4, pp. 656–664, 2002.

[14] S. Pei, R. Dong, R. L. He, and S. S.-T. Yau, "Large-scale genome comparison based on cumulative fourier power and phase spectra: central moment and covariance vector," *Computational and structural biotechnology journal*, vol. 17, pp. 982–994, 2019.

[15] A. Zielezinski, H. Z. Girgis, G. Bernard, C.-A. Leimeister, K. Tang, T. Dencker, A. K. Lau, S. Röhling, J. J. Choi, M. S. Waterman *et al.*, "Benchmarking of alignment-free sequence comparison methods," *Genome biology*, vol. 20, no. 1, pp. 1–18, 2019.

[16] K. Song, J. Ren, G. Reinert, M. Deng, M. S. Waterman, and F. Sun, "New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing," *Briefings in bioinformatics*, vol. 15, no. 3, pp. 343–353, 2014.

[17] J. Ren, X. Bai, Y. Y. Lu, K. Tang, Y. Wang, G. Reinert, and F. Sun, "Alignment-free sequence analysis and applications," *Annual Review of Biomedical Data Science*, vol. 1, p. 93, 2018.

[18] A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski, "Alignment-free sequence comparison: benefits, applications, and tools," *Genome biology*, vol. 18, no. 1, pp. 1–17, 2017.

[19] G. Mendizabal-Ruiz, I. Román-Godínez, S. Torres-Ramos, R. A. Salido-Ruiz, and J. A. Morales, "On dna numerical representations for genomic similarity computation," *PloS one*, vol. 12, no. 3, p. e0173288, 2017.

[20] E. Borrayo, E. G. Mendizabal-Ruiz, H. Vélez-Pérez, R. Romo-Vázquez, A. P. Mendizabal, and J. A. Morales, "Genomic signal processing methods for computation of alignment-free distances from dna sequences," *PloS one*, vol. 9, no. 11, p. e110954, 2014.

[21] B. Rim, N.-J. Sung, S. Min, and M. Hong, "Deep learning in physiological signal data: A survey," *Sensors*, vol. 20, no. 4, p. 969, 2020.