Classificação de genomas de vírus usando redes complexas e teoria da informação

Genome classification of viruses using complex networks and information theory

Gabriel Sporck Trombini[1], Matheus Henrique Pimenta-Zanon[2], Fabrício Martins Lopes[3]

RESUMO

A bioinformática desempenha um papel crucial na pesquisa e compreensão dos vírus, impulsionada pelo crescente e exponencial aumento das informações genômicas disponíveis. Este estudo adota uma abordagem, aproveitando as capacidades das redes neurais convolucionais (CNNs) e representando os dados sob a forma de matrizes de imagens, com o objetivo principal de simplificar a análise das complexas sequências de RNA. Essa transformação permite a aplicação de técnicasde processamento de imagens em ambientes genômicos, ampliando a compreensão das características subjacentes. Além disso, ao incorporar técnicas de redução de dimensionalidade nessa representação de imagens, a análise torna-se mais acessível e eficaz, proporcionando uma visualização maisclara e a interpretação de intricados padrões presentes em vastos conjuntos de dados genômicos. Embora os resultados iniciais deste estudo não tenham representado uma revolução significativa, é importante destacar que a integração da bioinformática e da inteligência artificial continua a desempenhar um papel fundamental na pesquisa biomédica em constante evolução.
PALAVRAS-CHAVE: Análise Genômica; Redes neurais convolucionais (CNNs); Sequências de RNA.

ABSTRACT

Bioinformatics plays a crucial role in the research and understanding of viruses, driven by the growing and exponential increase in available genomic information. This study adopts an approach, leveraging the capabilities of Convolutional Neural Networks (CNNs) and representing data in the form of image matrices, with the primary goal of simplifying the analysis of complex RNA sequences. This transformation allows the application of image processing techniques in genomic environments, expanding the understanding of underlying characteristics. Furthermore, by incorporating dimensionality reduction techniques in this image representation, the analysis becomes more accessible and effective, providing a clearer visualization and interpretation of intricate patterns present in vast genomic datasets. While the initial results of this study may not have represented a significant revolution, it is important to highlight that the integration of bioinformatics and artificial intelligence continues to play a fundamental role in the ever-evolving biomedical research.
KEYWORDS: Genomic Analysis; Convolutional Neural Networks; RNA Sequences.

INTRODUÇÃO

Nos últimos anos, avanços tecnológicos na área da genômica têm desempenhado um papel fundamental na geração de uma quantidade exponencial de dados genômicos. Através desses dados, os cientistas podem obter insights valiosos sobre a estrutura e

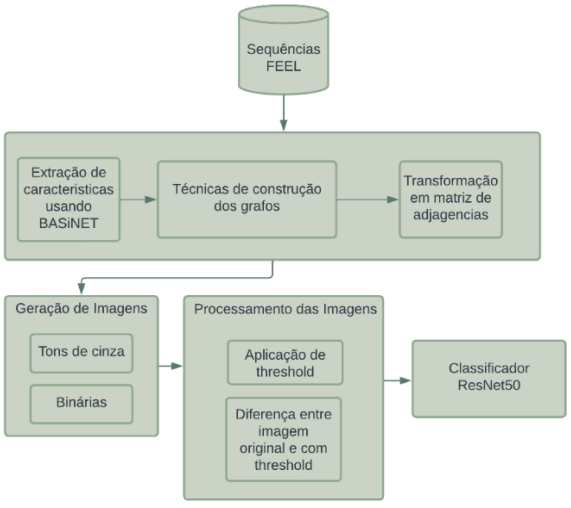
[1] Bolsista do(a) Araucária (Fundação Araucária de Apoio ao Desenvolvimento Científico e Tecnológico do Paraná). Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil. E-mail: gabrieltrombini@alunos.utfpr.edu.br. ID Lattes: 4016029853290740.
[2] Bolsista, Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil. E-mail: matheus.pimenta@outlook.com. ID Lattes: 4000443883671871.
[3] Docente no Curso de Engenharia de Computação. Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil. E-mail: fabricio@professores.utfpr.edu.br. ID Lattes: 1660070580824436.

função dos organismos vivos. No entanto, a análise dessas sequências biológicas complexas tem se tornado um desafio cada vez maior devido à sua crescente dimensão e complexidade. Um dos maiores problemas que esse estudo foca em ajudar é no alinhamento de sequências, o qual envolve a comparação de duas ou mais sequências de símbolos, como letras, aminoácidos ou bases nucleotídicas. Para identificar padrões de similaridade ou correspondência entre elas, o grande problema é que a quantidade de possíveis alinhamentos cresce exponencialmente com o tamanho das sequências envolvidas, fazendo com que o custo computacional seja inviável (FRANKE, 2021). Este trabalho se concentra em uma abordagem que visa simplificar a análise de sequências de RNA, adotando os princípios das redes neurais convolucionais (CNNs) e da representação de dados sob a forma de matrizes de imagens. A transformação de sequências biológicas em matrizes de imagens oferece uma nova perspectiva para lidar com essas informações, permitindo a aplicação de técnicas desenvolvidas para processamento de imagens em um contexto genômico. A hipótese central deste estudo é que ao representar sequências biológicas como imagens, torne possível capturar padrões estruturais em larga escala que podem não ser facilmente perceptíveis quando consideradas apenas as sequências de letras que compõem o DNA ou o RNA. Ao considerar as vantagens das CNNs, que foram projetadas para extrair características hierárquicas e espaciais de imagens, essa abordagem pode simplificar a detecção de informações relevantes nos dados genômicos.

METODOLOGIA

A metodologia adotada para a realização desta pesquisa de iniciação científica foi organizada em etapas distintas, a fim de alcançar os objetivos propostos. O diagrama abaixo ilustra a sequência de ações realizadas no decorrer do estudo:

Figura 1 – Arquitetura da pesquisa



Fonte: Autoria própria

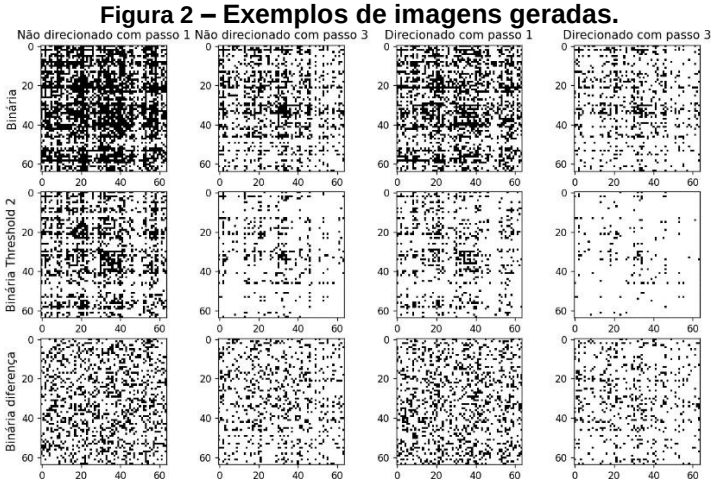
O Com isso as etapas de cada processo proposto são:

- **Coleta de Dados:** Para iniciar a pesquisa, foi adotado um conjunto de dados proveniente do trabalho FEELnc (WUCHER *et al.*, 2017), que consiste em um dataset de humanos, contendo duas classes distintas com 5 mil arquivos cada: RNA não codificante e RNA

mensageiro. Esse conjunto de dados foi selecionado estrategicamente para avaliar as técnicas propostas neste trabalho em um cenário desafiador. Compreender a eficácia das técnicas em um dataset mais complexo é crucial para garantir a robustez das análises realizadas. Posteriormente, empregamos as técnicas desenvolvidas nos conjuntos de dados das variantes do vírus de RNA (SARS-CoV-2), cada uma com mil arquivos: Alpha, Beta, Delta, Gamma e Omicron (PIMENTA-ZANON *et al*., 2022).

- **Extração de Característica:** O BASiNET (ITO *et al*., 2018) é uma ferramenta que não depende do alinhamento de sequências e é usada para classificar sequências biológicas. Ele faz isso ao extrair características a partir de medidas obtidas de redes complexas. Em suma ele funciona como uma "janela" que percorre ao longo da sequência de RNA com um tamanho específico. Em cada posição dessa janela, ele cria pontos que representam os valores dos nucleotídeos encontrados. Além disso, ele estabelece conexões entre esses pontos, seguindo a ordem em que os nucleotídeos aparecem na sequência de RNA. Essas conexões formam algo parecido com um mapa que mostra como os nucleotídeos estão relacionados entre si. É como se fossem marcadores em um caminho que ajuda a entender a ordem dos nucleotídeos na sequência de RNA. Um exemplo ilustrativo desse processo pode ser observado na Figura 2. Nesse contexto, uma sequência de RNA é analisada através de uma abordagem de janela deslizante, com um tamanho de janela definido como 3 (ITO *et al*., 2018).
- **Geração e Processamento das Imagens:** Transformação da matriz de adjacência em representações visuais assume um papel fundamental nesta pesquisa. Para realizar essa conversão, empregamos a linguagem de programação Python, fazendo uso de suas bibliotecas especializadas, permitindo a aplicação de diversas técnicas para otimizar os resultados obtidos. As técnicas exploradas neste estudo foram categorizadas em imagens binárias e tons de cinza. No caso das imagens binárias, o objetivo principal consistiu em destacar de forma clara e direta a presença ou ausência de conexões. Em contraste, para as imagens em tons de cinza, buscamos extrair significados dos valores contidos nas arestas, os quais poderiam conter informações valiosas. Entre as abordagens adotadas para ambas as categorias, incluiu-se a aplicação da técnica de *thresholding*. Essa técnica teve como intuito realçar as relações mais relevantes, considerando um limiar predefinido variando de 0 a 10, com incrementos de 2. Outra técnica empregada consistiu na subtração das imagens, obtendo-se a diferença entre a imagem na qual o *thresholding* foi aplicado e a imagem original. Essa abordagem visou destacar características ou regiões específicas das imagens, potencialmente facilitando o reconhecimento de padrões.

Figura 2 – Exemplos de imagens geradas.



Fonte: Autoria própria

- **Classificação das Imagens:** Ao término do processo, as imagens resultantes foram introduzidas nos algoritmos de classificação. Embora o foco central deste estudo repousasse mais nas técnicas de geração de imagens do que no próprio classificador, a etapa de classificação das imagens geradas detém relevância significativa na pesquisa. Com este propósito em mente, a escolha recaiu sobre a arquitetura ResNet-50, uma rede neural convolucional de renome, reconhecida por sua destreza na extração de características e por seu desempenho em tarefas de classificação de imagens. A ResNet-50 se destaca por suas camadas profundas, capazes de aprender de maneira intricada e hierárquica as características das imagens, tornando-a uma escolha pertinente para o nosso contexto. A avaliação do desempenho da ResNet-50 revelou-se crucial, pois essa análise nos permitiu discernir sua eficácia na classificação das imagens geradas a partir da matriz de adjacência. Parâmetros de avaliação como acurácia e entropia de perda (loss entropy) foram empregados para mensurar a performance do modelo (CUSSI; MACHACA ARCEDA, 2023)

RESULTADOS E DISCUSSÕES

Com a obtenção dos conjuntos de dados, deu-se início a toda a sequência de procedimentos conforme previamente elucidado. Através do uso do BASiNET (ITO *et al*., 2018), foi gerada a matriz de adjacências, e a partir dessa matriz, as imagens foram criadas aplicando-se diversas técnicas distintas. Para avaliar a eficácia das abordagens, foi empregada uma metodologia de validação cruzada. Esse método divide os dados em conjuntos de treinamento e teste, permitindo uma avaliação mais realista do desempenho dos modelos desenvolvidos. A avaliação das representações visuais geradas a partir das sequências foi conduzida por meio do modelo ResNet-50. Para a tarefa de classificação binária do primeiro conjunto de dados, foram adicionadas camadas apropriadas, incluindo a função de ativação Sigmoid na camada de saída para estimar a probabilidade da classe positiva, uma camada de achatamento (Flatten) para converter as saídas em um vetor unidimensional, e a função de perda de entropia cruzada binária (binary cross-entropy) para medir o erro durante o treinamento.

Uma das técnicas empregadas neste trabalho foi a utilização de grafos direcionados. Os grafos direcionados, também conhecidos como grafos orientados, são estruturas matemáticas que consistem em nós (ou vértices) interconectados por arestas direcionadas. Cada aresta possui uma direção, indicando a relação específica entre dois nós: um nó de origem e um nó de destino. Com a utilização de grafos direcionados neste trabalho, poderá permitir uma exploração mais profunda das relações e padrões presentes nos dados. Ao focalizar o dataset proveniente do FEELnc (WUCHER *et al.*, 2017) observamos que a disparidade entre as abordagens de imagens binárias e tons de cinza não se mostrou excessivamente acentuada, ficando em torno dos 71% de acurácia.

Nos experimentos realizados – no Passo 1 com grafo direcionado (1), no Passo 3 com grafo não direcionado (2) e direcionado (3) –, notou-se uma relativa uniformidade nos resultados dos experimentos (2) e (3), com taxas de acurácia próximas a 75% tanto para as representações binárias quanto tons de cinza. Uma observação notável foi a performance ligeiramente superior no caso do experimento (1), alcançando 78% de acurácia, embora ainda permaneça consideravelmente modesta.

Uma tentativa adicional consistiu em estender o número de épocas no experimento prévio que era de 10, no qual utilizamos o grafo direcionado no Passo 1. Essa abordagem resultou em uma melhoria modesta na taxa de assertividade, ultrapassando a marca de 80%. Isso também deu origem a uma nova preocupação: o fenômeno de *overfitting*.

Outras estratégias abordadas foi a de implementação do *threshold* e uma abordagem explorando a aplicação da diferença entre as imagens com *threshold* e as originais. Entretanto, os resultados obtidos não apresentaram variações de acurácia significativas, resultando em uma acurácia próxima a 70% em ambos casos.

Uma das possíveis abordagem para o problema do *overfitting* em trabalhos futuros é a adoção de técnicas de regularização. Isso pode incluir métodos como o *Dropout* e o *Weight Decay*, que foram desenvolvidos para controlar o crescimento excessivo dos parâmetros do modelo. Além disso, a aplicação de técnicas de normalização, como o *Batch Normalization* também pode ser benéfica. Essas abordagens têm o potencial de reduzir a variabilidade dos resultados nos dados de teste, tornando o modelo mais robusto e mais bem adaptado para generalizar para novos exemplos.

No que concerne aos conjuntos de dados das cinco variantes do SARS-CoV-2 (PIMENTA-ZANON *et al*., 2022), os resultados obtidos revelaram-se menos promissores. Ao explorar as técnicas propostas, não foi possível observar qualquer discrepância significativanos resultados, resultando em uma acurácia aproximada de apenas 20%. Essa baixa performance levanta questionamentos sobre a adaptabilidade dessas técnicas em cenáriosmais complexos e específicos, como no caso das variações do vírus em questão.

Essa disparidade substancial nos resultados entre os datasets do FEELnc (WUCHER *et al.*, 2017) e as variantes do SARS-CoV-2 (PIMENTA-ZANON *et al*., 2022) destaca a influência crucial do contexto do conjunto de dados sobre a eficácia das técnicas empregadas. A compreensão dessas variações pode fornecer insights valiosos para o desenvolvimento de abordagens mais robustas em análises futuras.

CONCLUSÃO

Apesar dos esforços e da abordagem inovadora empregada na pesquisa, os resultados obtidos ficaram aquém das expectativas. Ficou claro que a combinação da geração e processamento de imagens com a arquitetura ResNet-50 não alcançou o nível de desempenho desejado para a classificação das imagens geradas a partir da matriz de adjacência. Embora a solução proposta não tenha atingido os resultados desejados, os experimentos realizados para comparar essa abordagem com os métodos disponíveis na literatura forneceram insights valiosos sobre as limitações e desafios específicos dessa tarefa. É essencial reconhecer que a pesquisa científica frequentemente envolve desafios e descobertas imprevistas, e os resultados menos satisfatórios também contribuem para o avanço do conhecimento. Essa pesquisa pode servir como uma base sólida para futuros trabalhos, que podem explorar abordagens alternativas, refinamentos metodológicos e ajustes na escolha dos algoritmos, a fim de alcançar uma melhoria significativa na classificação de sequências genômicas de vírus.

Agradecimentos

Gostaríamos de expressar nossa profunda gratidão à Fundação Araucária, à Universidade Tecnológica Federal do Paraná (UTFPR) e ao Programa Institucional de Bolsas de Iniciação Científica (PIBIC) por seu apoio inestimável durante a realização deste projeto. A Fundação Araucária, por meio de seu generoso suporte financeiro, viabilizou os recursos necessários para a concretização deste trabalho. Seu compromisso com a pesquisa e o desenvolvimento acadêmico é uma pedra fundamental em nossa jornada.

Disponibilidade de código

A indisponibilidade do código-fonte se deve ao fato de que o projeto utiliza códigos de terceiros que não são de código aberto. Esses recursos não estão disponíveis publicamente devido a restrições legais e contratuais que proíbem a divulgação ou redistribuição do código-fonte. Portanto, não é possível disponibilizar o código desenvolvido neste contexto.

Conflito de interesse

Não há conflito de interesse.

REFERÊNCIAS

ITO, Eric Augusto; *et al*. **BASiNET—BiologicAl Sequences NETwork: a case study on coding and non-coding RNAs identification**. Nucleic Acids Research v. 46, 2018.

WUCHER Valentin *et al*. **FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome**. *Nucleic Acids Research*, v. 45, n. 8, p. e57, 2017.

FRANKE NEIA JUNIOR, Adilson. **Classificação de sequências de RNA utilizando redes neurais convulacionais**, Trabalho de Conclusão de Curso (TCC) - Engenharia de Computação, Universidade Tecnológica Federal do Paraná, 2021.

PIMENTA-ZANON, Matheus Henrique *et al*. **Biological Sequence Analysis Using Complex Networks and Entropy Maximization: A Case Study in SARS-CoV-2**. In: Swarnkar, T.; Patnaik, S.; Mitra, P.; Misra, S.; Mishra, M. (Eds.). Ambient Intelligence in Health Care. Smart Innovation, Systems and Technologies, vol. 317. Springer, Singapore, 2022.

CUSSI, Daniel Prado; MACHACA ARCEDA, V. E. **DNA Genome Classification with Machine Learning and Image Descriptors**. In: Future of Information and Communication Conference. Cham: Springer Nature Switzerland, p. 39-58, 2023.