ESTUDO DE IDENTIFICAÇÃO DE REGIÕES CODANTES EM GENES DE FUNGOS USANDO REDES NEURAIS CONVOLUCIONAIS

Álvaro de Araújo Ferreira Lima Neto (PIC/UEM), Josiane Melchior Pinheiro (Orientadora). E-mail: jmpferreira@uem.br.

Universidade Estadual de Maringá, Centro de Tecnologia, Maringá, PR.

Ciências exatas e da terra/Ciência da Computação

Palavras-chave: regiões de *splicing*; íntrons e éxons; aprendizado de máquina.

RESUMO

Este trabalho estuda a possível identificação de regiões codantes em genes de fungos do gênero *Colletotrichum* por meio do uso de Redes Neurais Convolucionais (CNNs). A pesquisa envolveu a coleta de sequências genéticas do *Genbank*, sendo aplicados filtros específicos para selecionar dados relevantes. As sequências foram transformadas em matrizes e utilizadas como entrada para a CNN. A metodologia incluiu a construção e treinamento do modelo, onde se buscava a classificação entre íntrons e éxons nas sequências genéticas. Os resultados mostraram que, apesar dos esforços para treinar o modelo, a acurácia obtida ainda foi consideravelmente baixa. A análise das métricas de precisão, *recall* e F1-*score* revelou que o modelo não conseguiu detectar padrões de forma eficaz. Uma possível limitação foi apontada na arquitetura da CNN, que não se adequou completamente à natureza do problema de classificação das sequências de DNA. Este estudo contribuiu para a compreensão da identificação de regiões codantes em genes de fungos do gênero *Colletotrichum* por meio de Redes Neurais Convolucionais. No entanto, os resultados indicam a necessidade de abordagens alternativas e ajustes na modelagem para alcançar melhores desempenhos.

INTRODUÇÃO

A evolução da bioinformática desde a descoberta e estudo dos genomas humanos resultou em progressos notáveis para a área, permitindo a manipulação eficiente de dados biológicos por meio de programas computacionais especializados (Lehugeur; Melo, 2018). Dentro desse contexto, o aprendizado de máquina desempenha um papel crucial na análise de grandes volumes de informações biológicas, abrangendo desde a previsão de estruturas proteicas até a modelagem de interações proteína-proteína (Dias; Pascutti; Silva, 2016).

A identificação precisa dos éxons, regiões codantes nos genes, é um desafio na análise do genoma. Muitas ferramentas baseiam-se em algoritmos de alinhamento, porém, repetições no genoma e erros de leitura tornam essa tarefa complexa (Chu; Li; Wu, 2015). O *splicing* de RNA é o processo de retirada dos íntrons (regiões não

codantes) e a ligação dos éxons remanescentes antes do processo de tradução do RNA para a proteína correspondente. Esse processo é fundamental para a regulação gênica pois permite a produção de proteínas diversas a partir da mesma sequência gênica, aumentando a complexidade celular.
Redes neurais têm sido usadas com sucesso para identificar regiões de *splicing*, e a aprendizagem profunda (que treina modelos de redes neurais) têm impactado várias áreas, incluindo a identificação de regiões de *splicing*, análises de sentimentos em textos e reconhecimento de padrões.
Este estudo explorou o uso de redes neurais convolucionais na identificação de regiões de *splicing* em sequências de RNA. Embora os resultados tenham sido limitados, o estudo destaca a valiosa lição de que nem sempre a pesquisa científica resulta em bons resultados imediatos, mas pode mostrar caminhos importantes para a ciência no futuro.

MATERIAIS E MÉTODOS

Para conduzir este estudo, diversos passos foram adotados, começando pela coleta de dados genéticos de fungos do gênero *Colletotrichum*. As sequências foram obtidas a partir do *GenBank*, um banco de dados de sequências genéticas mantido pelo NIH (Institutos Nacionais de Saúde dos EUA). A coleta foi orientada por critérios específicos, como o reino Fungi, o gênero *Colletotrichum* e o intervalo de datas desejado, além da marcação das regiões de íntron e éxon.
Em seguida, um código desenvolvido por Cruz e Menossi (2022) foi empregado para extrair informações essenciais das sequências, como nomes de espécies, proteínas associadas, sequências genéticas e localizações de íntrons e éxons. Os dados resultantes foram organizados e armazenados em um arquivo CSV para análises posteriores. Por se tratar de um problema de aprendizado supervisionado, este arquivo possui as sequências gênicas e as classificações de intervalos de íntrons e éxons correspondentes.
Foram coletadas as sequências inseridas no *GenBank* no período de janeiro a julho de 2022, obtendo 4890 sequências genéticas de 142 espécies de *Colletotrichum*. Dentre essas sequências, 896 contêm exclusivamente íntrons, 3994 possuem ambas as regiões e nenhuma é composta somente por éxons. As proteínas mais frequentes incluem a beta-tubulina, gliceraldeído-3-fosfato desidrogenase, quitina sintase e actina. A análise do tamanho médio das sequências indicou 466 códons, com mediana de 290 códons e foram identificadas seis bases degeneradas. Um códon é uma sequência de três nucleotídeos.
As sequências genéticas obtidas do *GenBank* possuem tamanhos variados (quantidade de bases nitrogenadas diferentes), enquanto a CNN precisa de uma entrada de tamanho fixo. Sendo assim, foi utilizada a técnica de janela deslizante, subdividindo as sequências de tamanhos variados, em sequências com 15 bases nitrogenadas. Após o processo de obter as sequências de tamanho fixo, elas foram transformadas em matrizes 15 x 4, sendo as linhas correspondentes às 15 bases nitrogenadas da sequência e as 4 colunas correspondentes aos nucleotídeos Adenina (A), Citosina (C), Guanina (G) e Timina (T), sendo o valor um para a base

representada e zero para as demais. As matrizes resultantes deste processo formam a base de dados de entrada para a CNN, que é treinada para identificar padrões nas matrizes e classificar as bases nitrogenadas como íntrons ou éxons. Para as bases degeneradas é feita a divisão do valor um pelas bases nitrogenadas possíveis que a base degenerada representa, de acordo com a União Internacional de Química Pura e Aplicada (IUPAC).

A CNN foi construída utilizando as bibliotecas *TensorFlow* e *Keras*. A arquitetura da rede compreendeu duas camadas convolucionais, cada uma com 15 filtros e um tamanho de kernel de (2, 2), seguidas por camadas de *MaxPooling* com tamanho de pool de (2, 2). Após as camadas convolucionais e de *pooling*, a rede possui uma camada *Flatten* para achatar os dados em um vetor unidimensional. Em seguida, há três camadas densas com 64, 32 e 15 neurônios, respectivamente. A última camada densa utiliza a função de ativação *softmax* para produzir probabilidades de classe de cada base nitrogenada (íntron ou éxon). A entrada da rede é composta por 15 bases nitrogenadas e a saída é uma probabilidade de classificação em íntron ou éxon para cada uma das bases nitrogenadas de entrada.

RESULTADOS E DISCUSSÃO

Este estudo analisou sequências genéticas do gênero *Colletotrichum* do *GenBank*. Os dados oriundos do *GenBank* foram transformados em matrizes 15 x 4 para entrada em uma CNN, que visava identificar íntrons e éxons. A base de dados de exemplos de matrizes, construída com base na técnica de janela deslizante, é composta por volta de 1.000.000 matrizes, sendo 18,33% delas de íntrons, 0% de éxons e 81,67% de misturas de íntrons e éxons.

Durante o treinamento da CNN, várias métricas de avaliação do modelo foram monitoradas, incluindo a área sob a curva (AUC), função de perda e acurácia. Embora a AUC tenha excedido 0,55 e a função de perda tenha diminuído, a acurácia permaneceu baixa. Métricas como precisão, *recall* e F1-*score* foram calculadas, indicando que o modelo não conseguiu detectar padrões de maneira satisfatória nos dados. A precisão do modelo foi de 0,69, o *recall* foi de 0,07 e o F1-*score* foi de 0,13. Esses resultados sugerem que a arquitetura da CNN com os parâmetros selecionados não capturou os padrões relevantes nas sequências, impedindo uma classificação eficaz.

Futuros estudos podem explorar abordagens alternativas, como a redução da saída da rede para categorizar as sequências de forma mais eficiente, além de examinar métricas alternativas e estratégias de pré-processamento mais sofisticadas.

Em resumo, este estudo ressalta a complexidade de identificar íntrons e éxons em sequências genéticas do gênero *Colletotrichum*. Embora os resultados não tenham alcançado os objetivos, as lições aprendidas e as oportunidades de aprimoramento são valiosas para pesquisas futuras nesta área.

CONCLUSÕES

Neste projeto, realizou-se a coleta e análise de sequências genéticas de fungos do gênero *Colletotrichum* do *GenBank*, utilizando critérios específicos. A análise revelou informações sobre proteínas predominantes, tamanho das sequências e bases nitrogenadas degeneradas. Para construir a base de dados, as sequências foram transformadas em matrizes 15 x 4 usando a técnica de janela deslizante. A base de dados foi composta por mais de 1.000.000 de exemplos de matrizes.

Um modelo de CNN foi utilizado para identificar íntrons e éxons tendo como entrada a base de dados de matrizes, mas infelizmente, os resultados não foram satisfatórios. As métricas de precisão, *recall* e F1-*score* não indicaram um bom desempenho. Mesmo com tentativas de ajustes após o término do projeto, os resultados permaneceram inalterados.

Apesar disso, o desenvolvimento deste trabalho sugere que futuros estudos possam explorar diferentes abordagens para modelar a base de dados, considerar outras métricas que sejam mais adequadas ao problema, entre outros. Além disso, os estudos sobre as redes neurais e os desafios de aplicá-las ao problema de identificação de regiões codantes, foram muito valiosos para o aprendizado dos autores.

REFERÊNCIAS

CHU, C.; LI, X.; WU, Y. Splicejumper: a classification-based approach for calling splicing junctions from rna-seq data. **BMC bioinformatics, BioMed Central**, v. 16, n. 17, p. S10, 2015.

DIAS, M. F. R.; PASCUTTI, P. G.; SILVA, M. L. da. Aprendizado de máquina e suas aplicações em bioinformática. ***Semioses***, v. 10, n. 1, p. 23–37, 2016.

LEHUGEUR, T. de P.; MELO, H. C. S. Bioinformática aplicada no desenvolvimento de novos fármacos**. *Psicologia e Saúde em debate***, v. 4, n. Suppl1, p. 55–55, 2018.

MENOSSI, V.; HENRIQUE FERREIRA CRUZ, G. **Um software para identificação de regiões codantes em genes de fungos filamentosos e sua tradução para os polipeptídeos correspondentes**. [s.l: s.n.].