XIII Seminário de Extensão e Inovação
XXVIII Seminário de Iniciação Científica e Tecnológica da UTFPR
Ciência e Tecnologia na era da Inteligência Artificial: Desdobramentos no Ensino Pesquisa e Extensão
20 a 23 de novembro de 2023 - *Campus Ponta Grossa, PR*
SEI-SICITE 2023

Histórias ocultas em BioDados: Análise de dados em bancos de dados biológicos

Hidden stories in BioData: Data analysis in biological databases

Yandriw Frederico Alicio de Lima[1], Alisson Gaspar Chiquitto[2], Alexandre Rossi Paschoal[3]

RESUMO

Para estabelecer um modelo de aprendizado de máquina que possa funcionar de maneira eficaz com novos dados, é fundamental realizar uma avaliação abrangente e uma comparação de vários conjuntos de dados antes do processo de treinamento. Esta etapa garante o desenvolvimento de modelos resilientes e flexíveis. A realização de análises exploratórias de dados nos *datasets* pode melhorar o reconhecimento das principais características e padrões pertinentes ao problema sob investigação, bem como identificar possíveis fontes de ruído e erros. Para enfrentar este desafio, este estudo conduziu uma pesquisa exploratória de análise de dados na comparação de vários conjuntos de dados relacionados a *long non-coding RNAs* (lncRNAs), RNA longos não codificantes. O objetivo foi destacar as distinções, ruídos e possíveis erros entre amostras da mesma espécie dos dados fornecidos por diferentes bancos de dados. Este estudo enfatizou a ideia de que nem todos os conjuntos de dados podem ser usados "brutos" em algoritmos de aprendizado de máquina. Em conclusão, este estudo forneceu *insights* valiosos sobre a combinação de pesquisa exploratória e seleção cuidadosa de conjuntos de dados para construir modelos de aprendizado de máquina precisos e confiáveis. Esta análise é de grande relevância pois enfatiza a importância de uma consideração cuidadosa da qualidade dos dados ao criar algoritmos de aprendizagem de máquina.
PALAVRAS-CHAVE: Análise Exploratória; Aprendizado de Máquina; lncRNA.

ABSTRACT

To establish a machine learning model that can work effectively with new data, it is critical to conduct a comprehensive assessment and comparison of multiple data sets prior to the training process. This step ensures the development of resilient and flexible models. Performing exploratory data analyzes on datasets can improve recognition of key features and patterns pertinent to the problem under investigation, as well as identify possible sources of noise and errors. To address this challenge, this study conducted exploratory data analysis research comparing various datasets related to *long non-coding RNAs* (lncRNAs). The aim was highlight distinctions, noise and possible errors between samples of the same species from data provided by different databases. This study emphasized the idea that not all datasets can be used "raw" in machine learning algorithms. In conclusion, this study provided valuable *insights* on combining exploratory research and careful selection of datasets to build accurate and reliable machine learning models. This analysis is of great relevance as it emphasizes the importance of careful consideration of data quality when creating machine learning algorithms.
KEYWORDS: Exploratory Data Analysis; Machine Learning; lncRNA.

[1] Bolsista do CNPq. Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil. E-mail: yandriw@alunos.utfpr.edu.br. ID Lattes: 7099116057503463.
[2] Doutorando pelo PROGRAMA DE PÓS-GRADUAÇÃO ASSOCIADO EM BIOINFORMATICA - PPGAB, Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil. E-mail: chiquitto@gmail.com. ID Lattes: 6062023075074006.
[3] Docente no Departamento Acadêmico de Computação. Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil. E-mail: paschoal@utfpr.edu.br. ID Lattes: 5834088144837137.

INTRODUÇÃO

Para desenvolver um modelo de aprendizado de máquina eficiente para processamento de dados novos, é de suma importância conduzir uma análise abrangente e comparativa de diversos conjuntos de dados antes do processo de treinamento. Isso garante a criação de modelos adaptáveis e resilientes. A realização de uma análise exploratória desses conjuntos de dados auxilia na identificação de características e padrões pertinentes ao problema em questão, além de possibilitar a identificação de possíveis fontes de imprecisões. Mediante a aplicação de conceitos fundamentais, como médias, medianas e representações gráficas, é possível detectar discrepâncias nos dados antes de empregá-los no treinamento do modelo. O pré-processamento dos dados, inclusive por meio de visualizações preliminares, contribui significativamente para o entendimento de conjuntos de dados relacionados à área da bioinformática.

Este trabalho desenvolveu uma pesquisa exploratória de análise de *datasets* relacionados a lncRNAs, RNAs longos não codificantes. O objetivo foi destacar as distinções, ruídos e possíveis erros entre amostras dos dados de sequências da mesma espécie fornecidos por diferentes *datasets*. Essa pesquisa buscou enfatizar que é necessário analisar os dados em primeira mão antes de treinar os algoritmos de aprendizado de máquina. Somado a isso levantou-se a hipótese de que as variações e erros encontrados nos *datasets*, gerariam impactos negativos nos modelos de *machine learning*.

METODOLOGIA

As sequências FASTA (nucleotídeos) de três bancos de dados de lncRNAS, sendo: o PLncDB V2.0 (JIN et al., 2020) que integra informações de recursos como EVLncRNAs e RNAcentral; o GreeNC 2.0 (DI MARSICO et al., 2021) e o CANTATAdb 2.0 (SZCZEŚNIAK et al., 2019) foram coletadas de seus sítios *web*.

Em seguida, foi feito o o pré-processamento dos bancos com a biblioteca Biopython (COCK et al., 2009). No caso, os arquivos do tipo FASTA foram convertidos em listas para facilitar a sua manipulação. As espécias *A. thaliana* e *O. sativa* foram selecionadas para o teste do modelo de *Machine Learning*, uma vez que eram as espécies em comum entre os bancos.

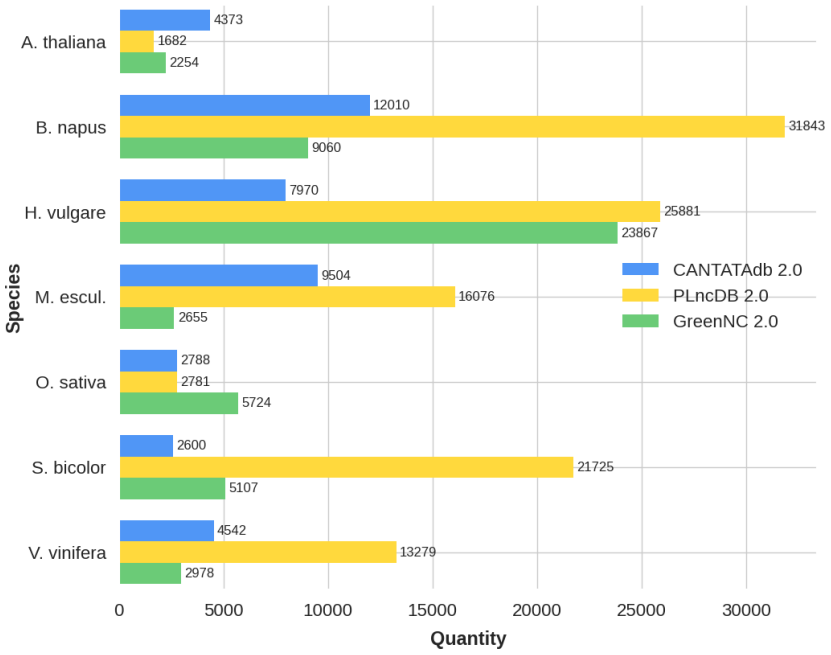
Para avaliar o potencial impacto das discrepâncias, ruídos e eventuais erros nas sequências provenientes dos bancos sobre o desempenho de um modelo de *Machine Learning*, foi desenvolvido um algoritmo de *Radom Forest* com a biblioteca Scikit-learn (PEDREGOSA et al., 2011) para realização de testes. Para evitar um *overfitting* foi utilizado um *cross validation* de valor 5. Para se testar esse algoritmo, foi criado um *dataset* de treino e um de teste com uma especificidade adicional. O *dataset* de treino possuía 20% das sequências de ambas as espécies dos *datasets* GreeNC 2.0 e o CANTATAdb 2.0, porém excluindo os dados do PlncDB 2.0. Já o *dataset* de teste continha os 80% restantes dos dados, juntamente com todo o banco PlncDB 2.0. Gerando assim um desafio adicional para o algoritmo de *Machine Learning*.

Toda a programação do projeto foi feita com a linguagem Python (PYTHON. . . , 2023) via uso do Google Colaboratory Notebook (GOOGLE, 2023). As ferramentas Matplotlib (HUNTER, 2007) e Seaborn (WASKOM, 2021) foram usadas para gerar as representações gráficas das análises.

RESULTADOS E DISCUSSÕES

Foi analisado as espécies em comum entre os bancos e suas respectivas quantidades de sequências de lncRNAs dos *datasets* (Figura 1). Vale ressaltar que a Figura 1 não apresenta toda a intersecção das espécies nos bancos de dados.

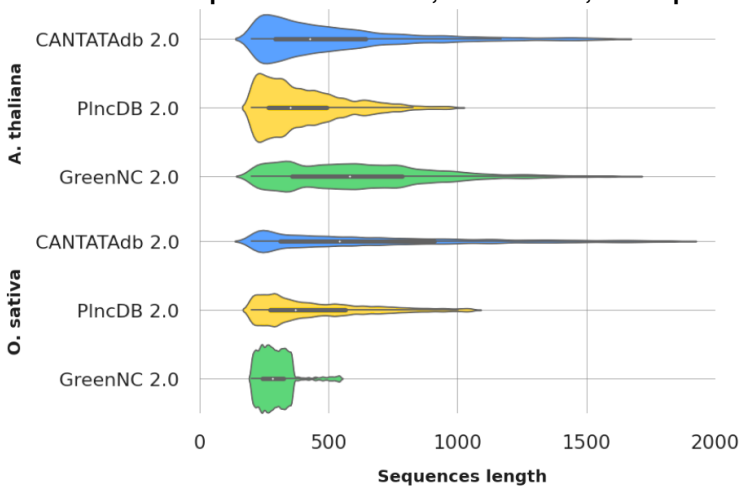
Figura 1 – Número de sequências de lncRNAs das espécies compartilhadas entre os bancos de dados



Fonte: AUTOR (2023).

A Figura 1, torna nítida a variação da quantidade de sequências entre os *datasets*, favorecendo a hipótese deste estudo. Já a Figura 2 representa graficamente os tamanhos das sequências, afim de se validar se a variação notada também se mantinha em outro aspecto dos dados.

Figura 2 – *Violinplot* do tamanho das seqências de lncRNAs, sem *outliers*, das espécies *A. thaliana* e *O. sativa*.

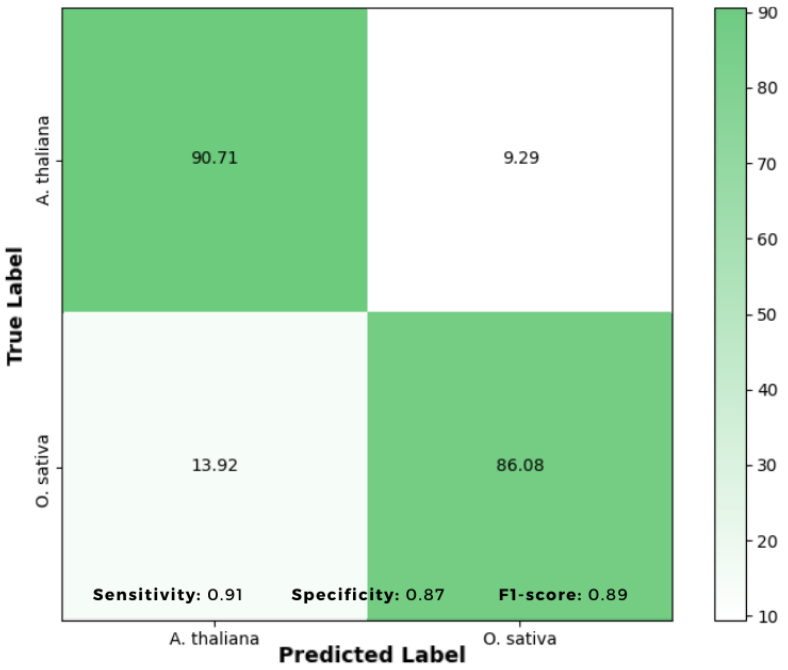


Fonte: AUTOR (2023).

Com base nos *violinplots* da Figura 2 foi possível notar a discrepância no tamanho das sequências disponibilizadas pelos *datasets*. No caso da espécie *O. sativa*, a mediana dos tamanhos das sequências do GreeNC 2.0 ficaram próximas do quartil inferior dos dados do PlncDB 2.0 e se localizaram ainda mais abaixo do quartil inferior do CANTATAdb 2.0. Logo a variação dos tamanhos fornecidos pelos *datasets* poderia gerar impactos no processo de *machine learning*.

O algoritmo de *Random Forest* foi executado para classificar as sequências como pertencentes a espécie *A. thaliana* ou *O. sativa*, pondo a prova a hipótese de que as inconsistências nos dados poderiam ser obstáculos para o modelo de *Machine Learning*. Os resultados desse experimento estão demonstrados na Figura 3.

Figura 3 – Matriz de confusão das espécies *A. thaliana* e *O. sativa*, juntamente com a sensibilidade, especificidade e *F1-score*.

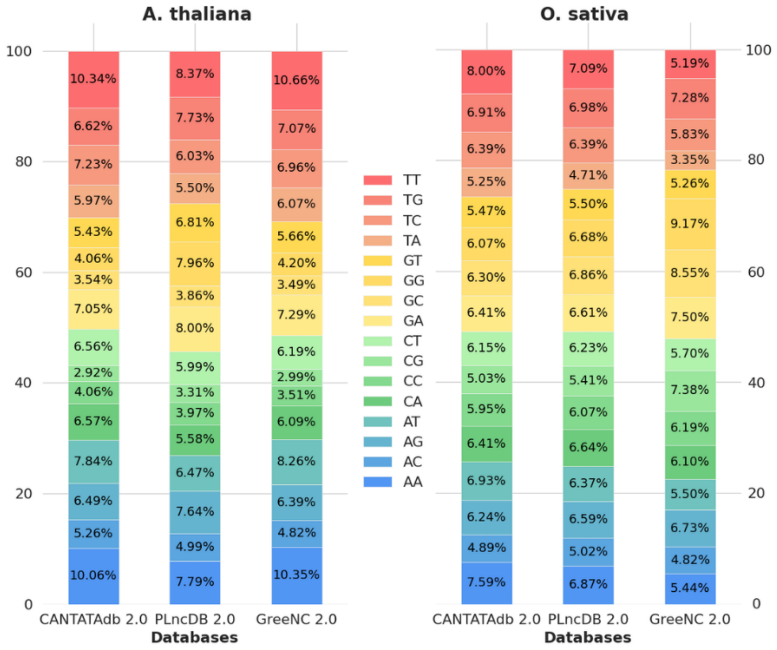


Fonte: AUTOR (2023).

Embora os tamanhos e números de sequências de lncRNAs variem amplamente entre os vários conjuntos de dados, os resultados obtidos foram satisfatórios. O *F1-score* de 0.89 demonstrou boa precisão do algoritmo. Logo, a hipótese de que tamanha inconsistência nos dados impactaria o processo de *machine learning* foi invalidada.

Analisando o problema com uma visão mais próxima a Biologia do que a Ciência de Dados, um fator mais intrínseco as sequências pôde ser explorado. A porcentagem de nucleotídeos e dinucleotídeos das sequências pode ter impactado os resultados. Quando analisadas tais porcentagens nas sequências de lncRNA ficou claro a semelhança dos dados disponibilizados pelos *datasets* apesar das variações de quantidade e tamanho. Isso é demonstrado pela Figura 4.

Figura 4 – Porcentagem de dinucleotídeos das espécies *A. thaliana* e *O. sativa*.



Fonte: AUTOR (2023).

CONCLUSÃO

Os *datasets* fornecerem uma gama variada de sequências e tamanhos de lncRNAs que, a princípio, não impactaram significativamente o desempenho do modelo de ML. Fundamentalmente o que foi o fator decisivo na classificação das espécies foram as porcentagens dos nucleotídeos e dinucleotídeos. Essa hipótese ainda pode ser mais bem avaliada em trabalhos futuros. Para garantir ainda mais relevância para este trabalho, futuramente algumas etapas do processo de coleta de dados e testes podem ser modificadas. Como por exemplo, validar se o uso exclusivo de banco de dados com sequências sintéticas ou provenientes de predições apresentaria resultados diferentes do que de *datasets* de dados reais. Em conclusão, este estudo demonstrou a combinação de pesquisa exploratória, visualização de dados e seleção cuidadosa de *datasets* para a construção de modelos de *machine learning* precisos e confiáveis.

CONFLITO DE INTERESSE

Não há conflito de interesse.

REFERÊNCIAS

COCK, Peter J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. **Bioinformatics**, v. 25, n. 11, p. 1422–1423, mar. 2009. ISSN 1367-4803. DOI: 10.1093/bioinformatics/btp163. eprint: https://academic.oup.com/bioinformatics/article-pdf/25/11/1422/48989335/bioinformatics_25_11_1422.pdf. Disponível em: .

DI MARSICO, Marco et al. GreeNC 2.0: a comprehensive database of plant long non-coding RNAs. **Nucleic Acids Research**, v. 50, n. D1, p. d1442–d1447, nov. 2021. ISSN 0305-1048. DOI: 10.1093/nar/gkab1014. eprint: https://academic.oup.com/nar/article-pdf/50/D1/D1442/42057523/gkab1014.pdf. Disponível em: .

GOOGLE. **Google Colaboratory**. [S.l.: s.n.], 2023. Disponível em: .

HUNTER, J. D. Matplotlib: A 2D graphics environment. **Computing in Science & Engineering**, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007. DOI: 10.1109/MCSE.2007.55.

JIN, Jingjing et al. PLncDB V2.0: a comprehensive encyclopedia of plant long noncoding RNAs. **Nucleic Acids Research**, v. 49, n. D1, p. d1489–d1495, out. 2020. ISSN 0305-1048. DOI: 10.1093/nar/gkaa910. eprint: https://academic.oup.com/nar/article-pdf/49/D1/D1489/35363903/gkaa910.pdf. Disponível em: .

PEDREGOSA, Fabian et al. Scikit-Learn: Machine Learning in Python. **J. Mach. Learn. Res.**, JMLR.org, v. 12, null, p. 2825–2830, nov. 2011. ISSN 1532-4435.

PYTHON. [S.l.]: Python.org, 2023. Disponível em: .

SZCZEŚNIAK, Michał Wojciech et al. CANTATAdb 2.0: Expanding the Collection of Plant Long Noncoding RNAs. In: **Plant Long Non-Coding RNAs: Methods and Protocols**. Edição: Julia A. Chekanova e Hsiao-Lin V. Wang. New York, NY: Springer New York, 2019. P. 415–429. ISBN 978-1-4939-9045-0. DOI: 10.1007/978-1-4939-9045-0_26. Disponível em: .

WASKOM, Michael L. seaborn: statistical data visualization. **Journal of Open Source Software**, The Open Journal, v. 6, n. 60, p. 3021, 2021. DOI: 10.21105/joss.03021. Disponível em: .