PATO: UMA FERRAMENTA UTILIZANDO MACHINE LEARNING PARA A PREDIÇÃO DE BACTÉRIAS PROBIÓTICAS

ISADORA COSENZA VIEIRA DA SILVA[1]; RAFAELLA SINNOTT DIAS[2]; FREDERICO SCHMITT KREMER[3]

[1]*Universidade Federal de Pelotas– isadoracosenza@gmail.com*
[2]*Universidade Federal de Pelotas – rafaellasinnott@gmail.com*
[3]*Universidade Federal de Pelotas – fred.s.kremer@gmail.com*

1. INTRODUÇÃO

Probióticos são definidos como microrganismos vivos, que, quando administrados em doses adequadas, conferem benefícios ao hospedeiro (WHO, 2001). Dentre estas bactérias, o gênero *Lactobacillus* é amplamente reconhecido pelo seu potencial probiótico e participação significativa na microbiota oral, vaginal e intestinal, mas seus mecanismos de ação ainda são pouco conhecidos (DIAS *et al*, 2022). Dessa forma, se faz importante a compreensão dos metabólitos primários e secundários produzidos a fim de entender os mecanismos de ação associados à ação probiótica (WIEERS *et al*, 2020).

A velocidade de obtenção de genomas foi drasticamente impactada pelo sequenciamento de nova geração, que permite a leitura de fragmentos de DNA de forma rápida e confiável (VAN DIJK *et al*, 2014). Por outro lado, a análise desses dados genômicos em larga escala requer ferramentas avançadas, como as oferecidas pela bioinformática. A bioinformática possibilita a análise e interpretação de vastas quantidades de dados biológicos, permitindo que pesquisadores identifiquem genes associados a características probióticas, analisem vias metabólicas e façam predições sobre a funcionalidade de microrganismos (UESAKA *et al*, 2022).

Recentemente, o uso de técnicas de *machine learning* tem se mostrado promissor na previsão de cepas probióticas, permitindo que computadores aprendam padrões a partir de grandes conjuntos de dados, facilitando a identificação de novas cepas probióticas com base em suas características genômicas e fenotípicas (MCCOUBREY *et al*, 2022). Apesar de existirem ferramentas que realizam a predição probióticos com base na sua genômica, como é o caso de *iProbiotics* (SUN *et al*, 2022), nenhuma analisa tanto aspectos genômicos quanto funcionais.

Portanto, a ferramenta *Probiotic Analysis Tool* (PATo) surge como uma solução inovadora nesse cenário, utilizando *machine learning* para a predição de probióticos com base em dados genômicos e funcionais. Essa abordagem tem o potencial de acelerar a descoberta de novas cepas probióticas e de esclarecer os mecanismos de ação que conferem benefícios à saúde do hospedeiro.

2. METODOLOGIA

Os genomas foram obtidos pela ferramenta online *iProbiotics*, conforme o modelo 3 (*Lactobacillus* Probiótico - *Lactobacillus* Não Probiótico). As sequências

genômicas e suas anotações foram recuperadas do NCBI usando as ferramentas NCBI Datasets (https://www.ncbi.nlm.nih.gov/datasets/) e Entrez-Direct. O número de genomas recuperados foram 58 *Lactobacillus* probiótico e 58 *Lactobacillus* não probiótico.

A anotação dos genomas foi feita usando *COGClassifier* (https://pypi.org/project/cogclassifier/), que atribui termos COG para identificar grupos funcionais gerais, e *AntiSMASH* (https://docs.antismash.secondarymetabolites.org/), que prevê genes associados a metabólitos secundários. As contagens geradas por ambas as ferramentas foram mescladas e normalizadas para representar a abundância relativa dos processos biológicos.

O pré-processamento de dados e o treinamento de modelos foram realizados com *Python* (https://www.python.org/), *Scikit-Learn* (https://scikit-learn.org/stable/), *Imbalanced-Learn* (https://imbalanced-learn.org/stable/), *BioPytho*n (https://biopython.org/) e *Pandas* (https://pandas.pydata.org/). Os dados foram balanceados com subamostragem aleatória e divididos em conjuntos de treino e teste. Modelos de machine learning foram treinados com validação cruzada (k=7) e avaliados com Acurácia, *Recall*, Precisão, F1-*Score* e ROC-AUC. Modelos com os maiores F1-*Score* e ROC-AUC foram analisados com SHAP (https://github.com/shap/shap) para entender a contribuição das características.

Por fim, uma aplicação web em *Python* foi desenvolvida com *Flask* (https://flask.palletsprojects.com/en/3.0.x/) e *Celery* (https://docs.celeryq.dev/en/stable/getting-started/introduction.html) para implantar o melhor modelo de cada conjunto de dados. Ela recebe genomas no formato *GenBank* e gera relatórios com anotações de *COGClassifier*, *AntiSMASH* e análise de características *SHAP* (https://github.com/shap/shap).

3. RESULTADOS E DISCUSSÃO

Neste trabalho serão discutidos os resultados do Modelo 3, que compara *Lactobacillus* probióticos e *Lactobacillus* não probióticos testados utilizando métricas de avaliação do modelo de classificação. As métricas foram calculadas usando a técnica de *holdout*, que consiste em dividir o conjunto de dados em duas partes: uma parte é usada para o treinamento do modelo e a outra parte para testar seu desempenho, permitindo avaliar como o modelo se comporta em dados que não foram previamente vistos. Nesse sentido, 75% dos genomas obtidos foram usados para treinamento do modelo, assertindo que o conjunto de dados tenha sido exposto a uma ampla gama de possibilidades, aumentando assim a acurácia da ferramenta.

Tabela 1: melhores resultados obtidos por *holdout* na classificação de *Lactobacillus*

Features	Accuracy	Recall	Precision	F1
funcional	94.74%	94.74%	94.74%	94.74%
5-mers	94.74%	94.74%	95.24%	94.72%
8-mers	94.74%	94.74%	95.24%	94.72%
9-mers	94.74%	94.74%	95.24%	94.72%
4-mers + funcional	92.11%	92.11%	93.18%	92.06%
5-mers + funcional	92.11%	92.11%	93.18%	92.06%
6-mers	92.11%	92.11%	93.18%	92.06%
3-mers	89.47%	89.47%	89.47%	89.47%
2-mers	89.47%	89.47%	89.92%	89.44%
3-mers + funcional	89.47%	89.47%	89.92%	89.44%
6-mers + funcional	89.47%	89.47%	89.92%	89.44%
8-mers + funcional	89.47%	89.47%	91.30%	89.36%

A Tabela 1 apresenta os 12 melhores resultados obtidos pela técnica de *holdout* para a classificação de *Lactobacillus* probióticos. Entre as características (*features*) utilizadas, as de melhor performance foram as características funcionais, seguidas de *k-mers* de 5, 8 e 9 nucleotídeos, respectivamente. Além dessas, outras combinações de *features* unindo *k-mers* e características funcionais também foram testadas, desempenhando similarmente. Os métodos de avaliação utilizados foram acurácia, *recall*, precisão e *F1 score*. Essas métricas foram essenciais para o treinamento correto dos modelos, pois proporcionaram uma avaliação detalhada e equilibrada do desempenho, permitindo ajustes e melhorias contínuas no processo de classificação. Assim, a sua combinação garante que o modelo seja eficaz em identificar e classificar corretamente as instâncias, resultando em maior confiabilidade.

Neste estudo, foi desenvolvida uma ferramenta de aprendizado de máquina para prever a atividade probiótica de microrganismos e explicar as características que determinam sua classificação como probióticos. Diferentemente do *iProbiotics*, que baseia-se apenas na composição de k-mers, PATo incorpora informações funcionais, proporcionando mais explicações sobre os mecanismos de ação dos probióticos. Enquanto *k-mers* são eficazes para classificações filogenéticas e de similaridade, eles não oferecem insights sobre funções biológicas ou vias metabólicas. Tanto a ferramenta quanto a aplicação web foram desenvolvidas com sucesso e podem ser acessadas em https://lnkd.in/dQkje_yA.

4. CONCLUSÕES

A ferramenta PATo se destaca pela sua abordagem inovadora ao analisar não apenas o genoma dos microrganismos, mas também suas características funcionais. Esta análise abrangente permite uma avaliação mais precisa da funcionalidade e mecanismos de ação desses probióticos. Dessa forma, PATo não só classifica os microrganismos com base em suas propriedades genômicas, mas também contribui significativamente para a compreensão dos processos que tornam esses microrganismos eficazes como probióticos.

5. REFERÊNCIAS BIBLIOGRÁFICAS

1. FAO/WHO. **Joint FAO/WHO consultation on evaluation of health and nutritional properties of probiotics in food including powder milk with live lactic acid bacteria.** Rome: World Health Organization, 2001.

2. DIAS, R.; KREMER, F.; AVILA, L. In silico prospection of Lactobacillus acidophilus strains with potential probiotic activity. **Braz J Microbiol**, Brasil, v. 54, n. 4, p. 2733-2743, 2023.

3. WIEËRS, G.; BELKHIR, L.; ENAUD, R.; LECLERCQ, S.; PHILIPPART DE FOY, J. M.; DEQUENNE, I.; DE TIMARY, P.; CANI, P. D. How Probiotics Affect the Microbiota. **Front Cell Infect Microbiol**, Estados Unidos, v. 9, p. 454, 2020.

4. VAN DIJK, E. L.; AUGER, H.; JASZCZYSZYN, Y.; THERMES, C. Ten years of next-generation sequencing technology. **Trends Genet**, Reino Unido, v. 30, n. 9, p. 418-426, 2014.

5. UESAKA, K.; OKA, H.; KATO, R.; KANIE, K.; KOJIMA, T.; TSUGAWA, H.; TODA, Y.; HORINOUCHI, T. Bioinformatics in bioscience and bioengineering: Recent advances, applications, and perspectives. **J Biosci Bioeng**, Japão, v. 134, n. 5, p. 363-373, 2022.

6. MCCOUBREY, L. E.; SEEGOBIN, N.; ELBADAWI, M.; HU, Y.; ORLU, M.; GAISFORD, S.; BASIT, A. W. Active machine learning for formulation of precision probiotics. **Int J Pharm**, Reino Unido, v. 616, p. 121568, 2022.

7. SUN, Y.; LI, H.; ZHENG, L.; LI, J.; HONG, Y.; LIANG, P.; KWOK, L. Y.; ZUO, Y.; ZHANG, W.; ZHANG, H. iProbiotics: a machine learning platform for rapid identification of probiotic properties from whole-genome primary sequences. **Brief Bioinform**, Reino Unido, v. 23, n. 1, p. bbab477, 2022.