Classificação de dados de metagenoma para detecção de vírus de interesse em saúde humana: monitoramento ambiental de SARS-Cov-2 como um caso de estudo

JEAN RODRIGUES OLIVEIRA DE SOUSA[1]; FREDERICO SCHMITT KREMER [3]

[1] *Universidade Federal de Pelotas – eujean.ros@gmail.com*
[3] *Universidade Federal de Pelotas – fred.s.kremer@gmail.com*

1. INTRODUÇÃO

O advento do sequenciamento de DNA de nova geração (*next generation sequencing*, NGS) permitiu um aumento expressivo na velocidade para obtenção de grandes volumes de dados genômicos, bem como de outras áreas das ciências ômicas, como transcriptômica e metagenômica (Liu et al., 2012). A metagenômica, também chamada "genômica ambiental", consiste na análise em larga de escala de dados derivados de sequenciamento de DNA de amostras complexas (metagenoma), que muitas vezes apresentam uma grande variedade de microorganismo. A análise do metagenoma pode ser realizada a partir da amplificação de marcadores taxonômicos seguida do sequenciamento ou a partir de técnicas de sequenciamento *shotgun* (*whole metagenome shotgun*), que permite também a caracterização funcional de genes (Bragg & Tyson, 2014).

Após a obtenção dos dados em um estudo de metagenoma é necessário se atribuir cada sequência identificada à uma determinada unidade taxonômica operacional (*operational taxonomic unit*, OTU), que pode ser um gênero, espécie ou mesmo sub-espécie que compartilha um determinado grau de similaridade de sequência (Mande et al., 2012). Este processo é denominado *binning*, e muitas das abordagens utilizam métodos de de alinhamento ou mapeamento de leituras como base (*alignment-based*), como BLAST (Altschul et al., 1990). De modo a reduzir o custo computacional, estratégias que não utilizam alinhamento de sequência (*alignment-free*). Estas estratégias podem ser utilizadas tanto para a classificação de amostras inteiras (ex: classificar o estado fisiológico de um indivíduo com base na composição da sua microbiota), como o realizado pela ferramenta DectICO (Ding et al., 2015), como classificar as leituras que pertencem a grupos taxonômicos de interesse, como a ferramenta Seeker (Auslander et al., 2020) e classificação de função de proteínas codificadas em metagenomas, como o Carnelian (Nazeen et al., 2020).

No contexto da pandemia de COVID19, Fongaro *et al* reportaram a presença de SARS-Cov2, agente etiológico da doença, em águas de esgoto coletadas no estado de Santa Catarina em Novembro de 2019, antes dos primeiros casos da doença serem oficialmente reportados na China, país onde a pandemia se iniciou (Fongaro et al., 2020). O monitoramento de água de esgoto para coronavirus, bem como outros doenças, vem sendo discutido em outros países, como os Estados Unidos, onde o *Center for Disease Control* (CDC) tem iniciado iniciativas como o *National Wastewater Surveillance System* (https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/wastewater-surveillance.html). A utilização de técnicas de metagenômica para a análise em larga escala de microorganismos presen-

tes em amostras de esgoto para fins de vigilância epidemiológica vem sendo discutida, sobretudo tendo em vista as possibilidades na detecção de vírus (Nieuwenhuijse et al., 2020) e genes de resistência à antibióticos (Hendriksen et al., 2019), possibilitando assim maior agilidade na detecção de surtos.

Neste contexto, no presente trabalho foi avaliada uma abordagem *alignment-free* baseada em vetorização de texto e aprendizado de máquina para classificação de dados derivados de análise metagenômica de modo a se identificar patógenos em amostras ambientais como as de monitoramento epidemiológico de esgoto, sendo usado como modelo de estudo o SARS-Cov-2.

2. METODOLOGIA

Dados de sequenciamento de viroma de esgoto foram obtidos do banco de dados *Sequence Read Archive* (SRA) do NCBI a partir do código de acesso, sendo o download realizado com a ferramenta `fastq-dump` do pacote SRA-Toolkit (https://github.com/ncbi/sra-tools). O genoma completo de SARS-Cov-2 foi obtido do Genbank a partir, sendo usado como base para a produção de leituras sintéticas similares à plataforma Illumina HiSeq com uso da ferramenta InSilicoSeq (https://github.com/HadrienG/InSilicoSeq). Ambos os conjuntos de dados foram utilizados como base para a produção de *subsets* de leituras positivas (simuladas) e negativas (de viroma de esgoto) tanto para treino quanto para validação do modelo preditivo. Além disso, um *subset* de leituras foi extraído dos dados de viroma ambiental para o treinamento de um modelo de vectorização de texto. Para o treinamento e validação dos modelos, em ambos os casos, foram utilizadas 10.000 leituras de cada classe. A vectorização de texto foi realizada com o algoritmo FastText (https://fasttext.cc/) que foi treinado a partir de uma representação das leituras de sequenciamento na forma de sequências de *k-mers* sobreponíveis (`k=11`, `vector_size=100`). Após o treinamento do modelo de vectorização, as leituras de treino e teste do modelo foram convertidas em sequências de k-mers e então vetorizadas. Então, um modelo de aprendizado de máquina baseado em *gradient-boosting* de árvores de decisão foi treinado usando a biblioteca XGBoost (https://xgboost.readthedocs.io/en/stable/). O modelo foi então avaliado utilizando as métricas acurácia e *recall* a partir do relatório de classificação gerado pela função `classification_report` do pacote Scikit-Learn.

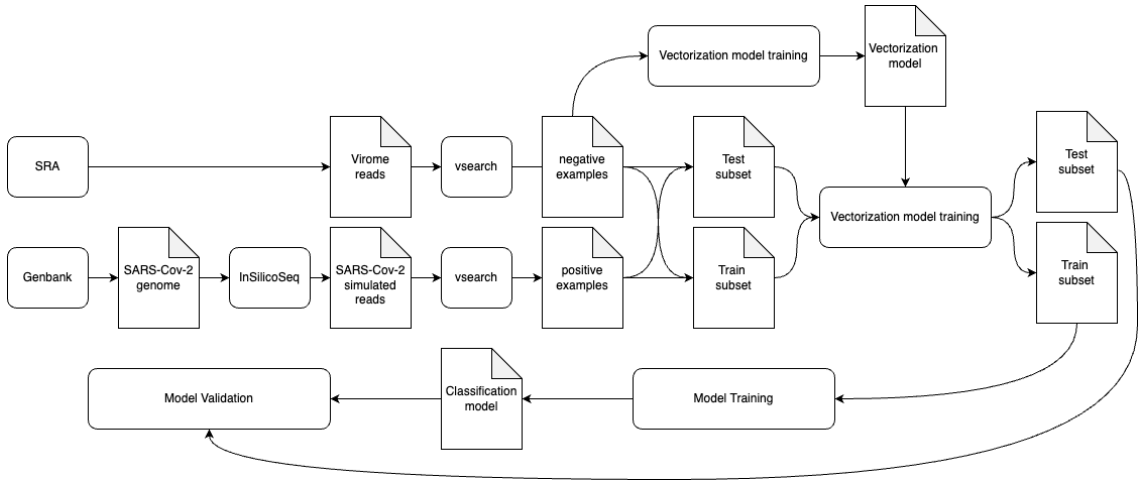

Figura 1. Resumo da metodologia adotada no presente trabalho.

3. RESULTADOS E DISCUSSÃO

A partir das representações geradas para as leituras de sequenciamento e do modelo de aprendizado de máquina produzido, foi possível se obter uma acurácia de 98% e um *recall* de 97% para a classe positiva na classificação, indicando com métodos de aprendizado de máquina podem ser uma alternativa na detecção de organismos patogênicos, como o vírus SARS-Cov-2, em amostras ambientais. Estes resultados vão de encontro com resultados de outros grupos, como Adjuik *et al* (2022), que demonstram que vetores de palavras podem ser usados para a classificação de proteínas deste vírus, mas estendem o escopo de aplicação para o uso de dados brutos de sequenciamento, sem necessidade de montagem ou anotação do genoma, tornando assim a detecção destes organismos mais ágil.

4. CONCLUSÕES

Com base na metodologia utilizada, foi possível demonstram que técnicas de vectorização de texto e aprendizado de máquina podem ser utilizadas para a produção de modelos preditivos capazes de auxiliar na detecção de organismos patogênicos de interesse. O código contendo as análises descritas neste trabalho está disponível no notebook do Google Colab e pode ser acessado através do link https://colab.research.google.com/drive/1VRpQ62AdtnBSWApAWPiC1npgBx16xvG?usp=sharing.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Adjuik, T. A & Ananey-Obiri, D. (2022). Word2vec neural model-based technique to generate protein vectors for combating COVID-19: a machine learning approach. *International Journal of Information Technology*.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410.

Auslander, N., Gussow, A. B., Benler, S., Wolf, Y. I., & Koonin, E. V. (2020). Seeker: alignment-free identification of bacteriophage genomes by deep learning. *Nucleic Acids Research*, 1.

Bragg, L., & Tyson, G. W. (2014). Metagenomics using next-generation sequencing. *Methods in Molecular Biology*, 1096, 183–201.

Ding, X., Cheng, F., Cao, C., & Sun, X. (2015). DectICO: An alignment-free supervised
metagenomic classification method based on feature extraction and dynamic selection. *BMC Bioinformatics*, 16(1), 323.

Fongaro, G., Stoco, P. H., Souza, D. S. M., Grisard, E. C., Magri, M. E., Rogovski, P., Schorner, M. A., Barazzetti, F. H., Christoff, A. P., Oliveira, L. F. V. de, Bazzo, M. L., Wagner, G.,

Hernandez, M., & Rodriguez-Lazaro, D. (2020). SARS-CoV-2 in human sewage in Santa
Catalina, Brazil, November 2019. *MedRxiv*, 2020.06.26.20140731.

Hendriksen, R. S., Munk, P., Njage, P., van Bunnik, B., McNally, L., Lukjancenko, O., Röder, T., Nieuwenhuijse, D., Pedersen, S. K., Kjeldgaard, J., Kaas, R. S., Clausen, P. T. L. C., Vogt, J.K., Leekitcharoenphon, P., van de Schans, M. G. M., Zuidema, T., de Roda Husman, A. M., Rasmussen, S., Petersen, B., ... Aarestrup, F. M. (2019). Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nature Communications*, 10(1), 1–12.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., & Law, M. (2012). Comparison of
next-generation sequencing systems. *Journal of Biomedicine & Biotechnology*, 2012, 251364.

Nazeen, S., Yu, Y. W., & Berger, B. (2020). Carnelian uncovers hidden functional patterns across diverse study populations from whole metagenome sequencing reads. *Genome Biology, 21*(1), 47.