APLICAÇÃO DE *EMBEDDINGS* NA CLASSIFICAÇÃO DE SEQUÊNCIAS VIRAIS EM DADOS METAGENÔMICOS DE MOSQUITOS DO GÊNERO AEDES

CARLOS A. C. S. JÚNIOR[1], JOÃO P. P. DE ALMEIDA[2], ULISSES B. CORRÊA[3]

[1] *Universidade Federal de Pelotas - cacsjunior@inf.ufpel.edu.br*
[2] *Universidade Federal de Minas Gerais - jpereiradealmeida.mg32@gmail.com*
[3] *Universidade Federal de Pelotas - ulisses@inf.ufpel.edu.br*

1. INTRODUÇÃO

Nos últimos anos, experienciamos diversas alterações climáticas significativas. A ocorrência de eventos extremos, aumento de temperatura e mudanças nos regime de chuvas propicia um ambiente favorável a proliferação de mosquitos principalmente do gênero *A. aegypti* e *A. albopictus* (WHITEHORN; YACOUB, 2019), conhecidos vetores de doenças como a Dengue, Zika e Chikungunya.

As arboviroses circulam entre animais silvestres, mantendo-se normalmente entre essas espécies. Com a proliferação desses mosquitos e o contato com novas espécies, provenientes do desmatamento, o ser humano acaba tornando-se um hospedeiro acidental. Apesar disso, uma quantidade pequena de vírus conseguem dar o salto de espécie, infectando e causando problemas ao hospedeiro não habitual (DONALISIO; FREITAS; ZUBEN, 2017).

Vírus possuem alta plasticidade genética e uma grande capacidade de mutação genética. Por serem necessariamente parasitas intracelulares, utilizam-se do maquinário do hospedeiro para sua replicação. Essa simplicidade, traz algumas desvantagens como a falta de elementos importantes para evitar erros na replicação. Essa propensão a erros e diversos outros fatores, contribuem para essa capacidade mutacional (MARKOV et al., 2023).

Estudos demonstraram que só em mosquitos do gênero *Aedes*, há uma diversidade de aproximadamente 80 espécies distintas de vírus (AGBOLI et al., 2019). Portanto é de grande relevância do ponto da saúde pública, identificar e analisar essas diversas espécies a fim de desenvolver políticas públicas para mitigar surto de casos como os da Dengue.

Para que essa investigação e caracterização seja possível, é necessário extrair o material genético desses vírus. Entretanto essa é uma tarefa complexa, uma vez que para isolar um vírus alvo, é necessário uma série de etapas de filtragem e cultivo em laboratório. Muitos não sobrevivem a essas etapas e impossibilitam seu estudo (WADELL, 1983).

Alternativamente, a área da metagenômica, trouxe diversos avanços na descoberta de novos vírus (SAKOWSKI et al., 2014). Apesar desses avanços, há uma série de desafios adicionais. Diferentemente da metodologia supracitada, todo o microbioma do mosquito é extraído. Dessa forma, temos o material genético de tudo que está presente, i.e., vírus, bactérias, fungos e o próprio DNA do mosquito (CASTRIGNANO; NAGASSE-SUGAHARA, 2015).

Para realizar a identificação e classificação de sequências virais e não virais, tipicamente utiliza-se algoritmos de alinhamento local, tais como o BLAST (*Basic Local Alignment Search Tool*) (ALTSCHUL et al., 1990), onde uma sequência de nucleotídeos é extraída e comparada com todos presentes em um banco de dados, a fim de encontrar o mais similar e classificá-lo.

Entrentanto, essa abordagem depende de classificações anteriores já consolidadas e não abrange satisfatoriamente a alta capacidade mutacional dos vírus.

Estudos mostram que o sequênciamento de metagnômico em larga escala há uma prevalência de sequências desconhecidas, ou seja, que não obtiveram similaridade com nada presente nos bancos de dados, mas que podem ser vírus. (AGUIAR et al., 2015)

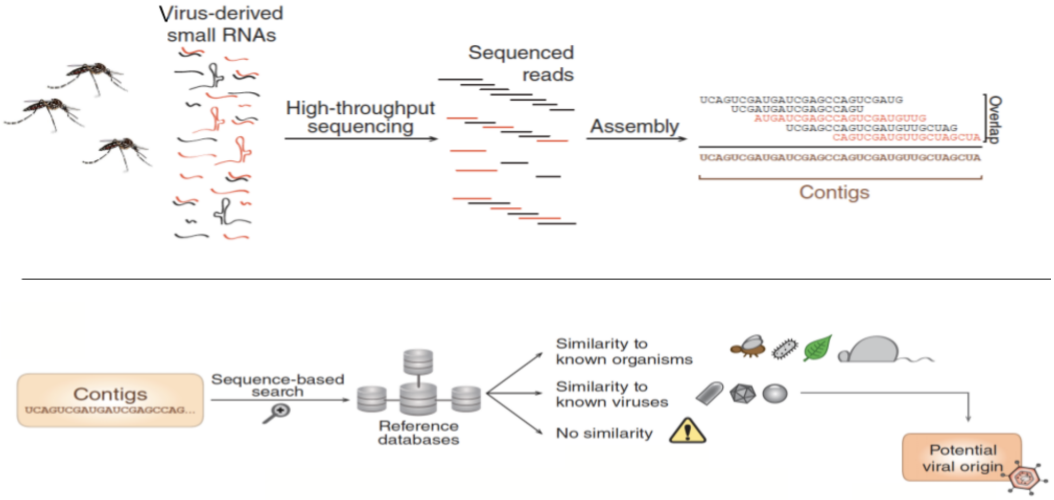
Portanto, o presente trabalho busca classificar as sequências virais em dados metagenômicos de mosquitos, utilizando o método de sequênciamento de larga escala e *shotgun*. A partir dos nucleotídeos recuperados, representaremos como string, a fim de extrair características latentes e ser capaz de realizar a classificação, utilizando *Embeddings* e classificadores.

2. METODOLOGIA

O processo de sequenciamento em larga escala é realizada em algumas etapas. Há diversos processos de sequência reconhecidos, entretanto iremos utilizar a apenas a abordagem *shotgun*. Nessa metodologia, em alusão ao tiro de escopeta, os genomas coletados são quebrados em pequenos fragmentos, o que é conhecido como *reads*. Estes são sequenciados através de uma máquina de sequência em larga escala *High-throughput sequencing*. Gera-se então grandes sequências de *reads*, que passam por um processo de filtragem e validação com ferramentas consolidadas para essa finalidade.

A partir dos redes gerados no processo anterior, é necessário montá-los para que contenham a sequência original completa. Isto é feito utilizando os fragmentos e através da sobreposição de sequência similares e contíguas, é realizada a montagem, conforme demonstrado na figura 1, obtendo assim os *contigs*. Ao final, esses *contigs* são alinhados com o banco de dados e fornecerá a classificação dessa sequência.

Figura 1: Pipeline descrevendo o processo de sequenciamento através do *High-throughput sequencing*



Fonte: (AGUIAR; OLMO; MARQUES, 2016)

O trabalho se insere neste ponto, onde a montagem desses *contigs* e seu alinhamento, resultam em genomas desconhecidos, e.g., genomas incompletos ou até um novo organismo ainda não conhecido.

Iremos utilizar dados públicos presentes no GenBank com sequências classificadas como vírus para treinar um modelo word2vec, a fim de capturar características latentes das sequências e gerar essa representação intermediária. Em seguida, utilizaremos os dados recentes de (OLMO et al., 2023), onde há um dataset novo de dados metagenômicos de mosquistos do gênero *Aedes*, já curados pela equipe e utilizá-los para avaliação de um modelo classificador de sequências virais.

Na tabela 1 é possível verificar a grande quantidade de dados sem alinhamento frente aos dados que foram classificados corretamente.

Tabela 1: Dataset

Origem	Virais	Não Virais	Sem Alinhamento
Dados Públicos	974	3814	127027
Pequenos RNAs	2315	10973	14673

3. PRÓXIMOS PASSOS E RESULTADOS ESPERADOS

Como parte de dissertação em progresso, ainda é necessário realizar os testes a avaliação correta dos modelos propostos a fim de validar sua eficiência.

No trabalho predecessor (FREITAS, 2020), utilizou-se redes convolucionais, adptadas para texto, para extrair as características das sequências e classificar corretamente. Essa abordagem gerou diversos desafios e problemas, como a restrição de entrada dos contigs, sendo necessário preenche-los quando maiores do que o definido.

Neste trabalho, espera-se melhorar os resultados utilizando o *embeddings* com *word2vec* e a utilização da expansão dos dados de pequenos RNAs, publicados após o trabalho supracitado.

4. REFERÊNCIAS

AGBOLI, E.; LEGGEWIE, M.; ALTINLI, M.; SCHNETTLER, E. Mosquito-specific viruses-transmission and interaction. **Viruses**, MDPI AG, v. 11, n. 9, p. 873, set. 2019.

AGUIAR, E. R. G. R.; OLMO, R. P.; MARQUES, J. T. Virus-derived small RNAs: molecular footprints of host–pathogen interactions. **Wiley Interdiscip. Rev. RNA**, Wiley, v. 7, n. 6, p. 824–837, nov. 2016.

AGUIAR, E. R. G. R.; OLMO, R. P.; PARO, S.; FERREIRA, F. V.; FARIA, I. J. d. S. de; TODJRO, Y. M. H.; LOBO, F. P.; KROON, E. G.; MEIGNIN, C.; GATHERER, D.; IMLER, J.-L.; MARQUES, J. T. Sequence-independent characterization of viruses based on the pattern of viral small RNAs produced by the host. **Nucleic Acids Res.**, Oxford University Press (OUP), v. 43, n. 13, p. 6191–6206, jul. 2015.

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215, n. 3, p. 403–410, 1990. ISSN 0022-2836. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0022283605803602>.

CASTRIGNANO, S. B.; NAGASSE-SUGAHARA, T. K. The metagenomic approach and causality in virology. **Revista de Saúde Pública**, Faculdade de Saúde Pública da Universidade de São Paulo, v. 49, 2015. ISSN 0034-8910. Disponível em: <https://doi.org/10.1590/S0034-8910.2015049005475>.

DONALISIO, M. R.; FREITAS, A. R. R.; ZUBEN, A. P. B. V. Arboviruses emerging in brazil: challenges for clinic and implications for public health. **Revista de Saúde Pública**, FAPESP, v. 51, p. 30, 2017. Acessado em: 9 Setembro 2024, Epub 10 Abr 2017, ISSN 1518-8787. Disponível em: <https://doi.org/10.1590/S1518-8787.2017051006889>.

FREITAS, L. C. **Classificação de sequências virais em dados de sequenciamento metagenômico de mosquitos vetores de arbovírus através de aprendizado de máquina**. 55 p. Dissertação (Trabalho de Conclusão de Curso) — Universidade Federal de Pelotas, Pelotas, 2020.

MARKOV, P. V.; GHAFARI, M.; BEER, M.; LYTHGOE, K.; SIMMONDS, P.; STILIANAKIS, N. I.; KATZOURAKIS, A. The evolution of sars-cov-2. **Nature Reviews Microbiology**, v. 21, n. 6, p. 361–379, Jun 2023. ISSN 1740-1534. Disponível em: <https://doi.org/10.1038/s41579-023-00878-2>.

OLMO, R. P.; TODJRO, Y. M. H.; AGUIAR, E. R. G. R.; ALMEIDA, J. P. P. de; FERREIRA, F. V.; ARMACHE, J. N.; FARIA, I. J. S. de; FERREIRA, A. G. A.; AMADOU, S. C. G.; SILVA, A. T. S.; SOUZA, K. P. R. de; VILELA, A. P. P.; BABARIT, A.; TAN, C. H.; DIALLO, M.; GAYE, A.; PAUPY, C.; OBAME-NKOGHE, J.; VISSER, T. M.; KOENRAADT, C. J. M.; WONGSOKARIJO, M. A.; CRUZ, A. L. C.; PRIETO, M. T.; PARRA, M. C. P.; NOGUEIRA, M. L.; AVELINO-SILVA, V.; MOTA, R. N.; BORGES, M. A. Z.; DRUMOND, B. P.; KROON, E. G.; RECKER, M.; SEDDA, L.; MAROIS, E.; IMLER, J.-L.; MARQUES, J. T. Mosquito vector competence for dengue is modulated by insect-specific viruses. **Nature Microbiology**, v. 8, n. 1, p. 135–149, Jan 2023. ISSN 2058-5276. Disponível em: <https://doi.org/10.1038/s41564-022-01289-4>.

SAKOWSKI, E. G.; MUNSELL, E. V.; HYATT, M.; KRESS, W.; WILLIAMSON, S. J.; NASKO, D. J.; POLSON, S. W.; WOMMACK, K. E. Ribonucleotide reductases reveal novel viral diversity and predict biological and ecological features of unknown marine viruses. **Proceedings of the National Academy of Sciences**, v. 111, n. 44, p. 15786–15791, 2014. Disponível em: <https://www.pnas.org/doi/abs/10.1073/pnas.1401322111>.

WADELL, G. Cultivation of viruses. In: **Textbook of Medical Virology**. [S.l.]: Elsevier, 1983. p. 38–44.

WHITEHORN, J.; YACOUB, S. Global warming and arboviral infections. **Clin Med (Lond)**, England, v. 19, n. 2, p. 149–152, mar. 2019.