Mauricio Souza Menezes

Ferramenta Computacional para Estudo da Evolução de Espécies Virais Baseado no Uso de Códons

Salvador-BA

28 de junho de 2023

Mauricio Souza Menezes

Ferramenta Computacional para Estudo da Evolução de Espécies Virais Baseado no Uso de Códons

Áreas da Computação:
Bioinformática

Orientador:
PhD Diego Gervasio Frias Suárez

Coorientador:
PhD Vagner Fonseca

UNEB - Universidade do Estado da Bahia
Departamento de Ciências Exatas e da Terra
Colegiado de Sistemas de Informação

Salvador-BA

28 de junho de 2023

Folha de Aprovação

Anteprojeto sob o título provisório *Ferramenta Computacional para Estudo da Evolução de Espécies Virais Baseado no Uso de Códons* apresentado como exigência parcial para avaliação na disciplina Trabalho de Conclusão de Curso I do bacharelado em Sistemas de Informação da Universidade do Estado da Bahia entregue por *Mauricio Souza Menezes* a Maria Cristina Elyote Santos professora da disciplina, em 28 de junho de 2023, em Salvador, Bahia.

Mauricio Souza Menezes
Orientando

Diego Gervasio Frias Suárez

Sumário

1 Introdução

Os desafios impostos pela pandemia do COVID-19 incluíram a falta de conhecimento suficiente para a compreensão da importância das ameaças biológicas e para a preparação médica, apesar dos avanços científicos e tecnológicos. O conhecimento prévio sobre os agentes biológicos com potencial para causar pandemias pode melhorar substancialmente nossa preparação pré-pandemia. (1, p. 1)

Diante disso, a bioinformática, que é a junção de métodos computacionais e técnicas estatísticas com o objetivo de extrair informações de dados biológicos brutos, desempenha um papel fundamental na interpretação de dados genômicos e na compreensão de processos evolutivos. Segundo (2, p.1) "os métodos filogenéticos podem ser usados para analisar os dados da sequência de nucleotídeos de forma que a ordem de descendência de cepas relacionadas possa ser determinada. Quando associada à análise filogenética apropriada, a epidemiologia molecular tem o potencial de elucidar os mecanismos que levam a surtos microbianos e epidemias."

A reconstrução filogenética é uma das abordagens amplamente utilizadas na análise da evolução de espécies, que permite investigar as relações evolutivas entre diferentes linhagens de vírus. Essas observações são realizadas com base em dados como sequências de Ácido Desoxirribonucleico (DNA) e Ácido Ribonucleico (RNA). Essas sequências são formadas por blocos fundamentais chamados de nucleotídeos, que são compostos por uma base nitrogenada, um açúcar e um grupo fosfato. As bases presentes nos nucleotídeos do DNA são adenina (A), timina (T), citosina (C) e guanina (G), enquanto no RNA a base timina é substituída pela uracila (U). (3)

Uma das principais formas de análise filogenética é realizada através da árvore filogenética, onde são representadas as relações evolutivas entre um conjunto de espécies. De acordo com (4) elas tem função importante porque apresentam de forma sucinta e particular a evolução dos descendentes partindo de ancestrais em comum.

A semelhança genética entre vários vírus infecciosos e mortais fornece uma visão

do fato de que o RNA é a chave para discernir e marcar os possíveis patógenos que podem causar uma pandemia. Embora um padrão geral e motivos conservados possam ser observados em ancestrais imediatos, as regiões não conservadas das sequências são o resultado da acumulação de mutações, seja por inserção ou deleção de um ou vários nucleotídeos ou por substituição pontual de um nucleotídeo por outro. A fonte principal de mutações em vírus são percalços na replicação e a recombinação de RNA (1, p. 11).

Apesar da utilidade da filogenética e dos softwares comerciais e públicos disponíveis para análises filogenéticas, os métodos filogenéticos são muitas vezes aplicados de forma inadequada. Mesmo quando aplicados adequadamente, são mal explicados e, portanto, mal compreendidos. (2, p. 1) Além disso, por trabalhar com grandes quantidades de dados, os métodos utilizados devem ser avaliados também em relação ao seu custo computacional.

Na busca de trabalhos relacionados, vários métodos foram encontrados, e a seguir são apresentados.

O método de Máxima Verossimilhança (ML) (ou *Maximum Likelihood*), não é exclusivo da filogenia, mas sim uma abordagem estatística. A sua aplicação em filogenia consiste em avaliar a probabilidade de que o modelo de evolução escolhido gere os dados observados, que são por exemplo, características de um organismo. Essa proposta foi utilizada por (1, 5–11).

Já em (12, 13), foi usada a inferência bayesiana, que é fundamentada no teorema de Bayes, que permite a atualização das probabilidades a priori para probabilidades a posteriori à medida que novas evidências são incorporadas.

Além desses, (14) utilizou a junção de vizinhos (ou *Neighbor-Joining*), que é baseado em uma abordagem heurística que visa construir uma árvore filogenética a partir de uma matriz de distância entre as sequências estudadas. O trabalho de (15) expos o Frequency Chaos Game Representation e (16) a floresta aleatória. Por fim, (17) comparou três modelos para reconstrução de árvores filogenéticas: junção de vizinhos; ML e inferência bayesiana.

As soluções até então desenvolvidas, são guiadas pela reconstrução das árvores filogenéticas construídas a partir das mutações de nucleotídeos. Neste aspecto, as ferramentas disponíveis não oferecem uma aplicação no contexto de árvores reconstruídas com distâncias obtidas a partir da diferença do uso de códons, que são sequências de três nucleotídeos responsáveis pela codificação dos aminoácidos nas proteínas. Os códons desempenham um papel crucial na determinação da função e estrutura das proteínas, e

alterações nos códons podem resultar em mudanças significativas nas características fenotípicas dos vírus. Necessita-se então, de pesquisas e desenvolvimento de ferramentas que realizem uma classificação de sequências genéticas com base no uso/frequência de códons.

2 Objetivos

Com base no problema de pesquisa apresentado na seção anterior destacamos os seguintes objetivos a serem atingidos ao final da pesquisa.

2.1 Objetivo Geral

Desenvolver e validar um novo método de análise da evolução molecular viral

2.2 Objetivos Específicos

(i) Definir um modelo para validação do método proposto.

(ii) Desenvolver uma de ferramenta para caracterizar/validar o método.

(iii) Coletar os dados necessários para aplicar na ratificação do método.

(iv) Disponibilizar o modelo como uma ferramenta web de fácil acesso.

3 Justificativas e Contribuições

Desenvolver um método de construção de árvores com base nas distâncias obtidas a partir da diferença do uso de códons contribuiria com a tarefa de classificação de cepas para a vigilância sanitária, especialmente na descoberta de novas cepas emergentes com potenciais pandêmicos. Ademais, é também importante dispor de alternativas à filogenia molecular atualmente utilizada, para gerar informações de outro ponto de vista e-ou para servir de referência aos métodos filogenéticos. Os métodos atuais ainda demandam de um alto custo computacional, sendo assim, existe a necessidade de desenvolver outros mais baratos e que possam suportar o volume crescente de dados (sequências). Sendo assim, o projeto visa apresentar um método que seja capaz de realizar classificações, com um custo computacional baixo, em relação a outros métodos, e que possa apresentar, do ponto de vista científico, alternativas de comparação com outras técnicas já existentes.

4 Metodologia

Um ponto importante para a obtenção dos objetivos deste trabalho está relacionada a definição da metodologia que servirá como alicerce. Com a proposta de desenvolver e validar um método de análise da evolução molecular de vírus com base no uso de códons, a metodologia escolhida para isso é o Design Science Research (DSR). Essa metodologia, proporciona um framework teórico e prático para a criação de artefatos inovadores, como métodos, modelos ou frameworks, visando resolver problemas específicos.(18)

Para a obtenção de sucesso ao utilizar o DSR os seguintes passos serão seguidos:

1. Identificação do problema e definição dos objetivos.
2. Desenvolvimento do artefatos.
3. Avaliação do artefato.
4. Apresentar contribuições científicas.

Também será utilizada analises quantitativas, ou seja, medidas estatísticas para mensurar e comparar os resultados obtidos.
A pesquisa quantitativa só tem sentido quando há um problema muito bem definido e há informação e teoria a respeito do objeto de conhecimento, entendido aqui como o foco da pesquisa e/ou aquilo que se quer estudar. Esclarecendo mais, só se faz pesquisa de natureza quantitativa quando se conhece as qualidades e se tem controle do que se vai pesquisar.(19)

Os pontos a seguir serão realizados durante o desenvolvimento do projeto:

- Coleta de sequências que serão utilizadas.
- Tratamento necessário dos dados.
- Avaliação de desempenho do modelo.

- Analise comparativa com modelos convencionais.
- Disponibilização do modelo como ferramenta web.

5 Cronograma

As atividades descritas nas tabelas abaixo, correspondem aos processos realizados para alcançar os objetivos propostos na segunda seção deste trabalho. A tabela 1 apresenta as atividades realizadas seguindo o escopo da disciplina de Trabalho de Conclusão de Curso I, já a tabela 2 corresponde as atividades com conclusão previstas para a disciplina de Trabalho de Conclusão de Curso II.
Cada mês corresponde a aproximadamente 30(trinta) dias. São alocadas, pelo menos, 4(quatro) horas diárias para realização das atividades.

	Jan	Fev	Mar	Abr	Mai	Jun
Definição dos objetivos e questões de pesquisa	•					
Revisão sistemática da literatura e descrição do trabalho em forma de relatório		•	•			
Elaboração deste texto do anteprojeto				•	•	
Formulação da hipótese identificada na pesquisa, que resultará no método proposto para solução					•	•
Apresentação da hipótese da pesquisa sobre um novo método de caracterização genômica						•

Tabela 1: Cronograma de Janeiro até Junho de 2023

	Jul	Ago	Set	Out	Nov	Dez
Desenvolvimento da primeira versão do método	•	•	•			
Montagem dos datasets	•	•	•			
Avaliação formativa do método de acordo com a metodologia proposta para o projeto			•	•		
Definir e implementar a ferramenta web			•	•	•	
Testar e implementar as melhorias necessárias para atingir o objetivo				•	•	
Análise e considerações do resultado final					•	
Elaboração e escrita da monografia	•	•	•	•	•	•

Tabela 2: Cronograma de Julho a Dezembro de 2023

Referências

1 BEHL, A. et al. Threat, challenges, and preparedness for future pandemics: A descriptive review of phylogenetic analysis based predictions. *Infection, Genetics and Evolution*, v. 98, p. 105217, mar. 2022. ISSN 15671348. Disponível em: ¡⟨https://linkinghub.elsevier.com/retrieve/pii/S1567134822000144⟩¿.

2 HALL, B. G.; BARLOW, M. Phylogenetic analysis as a tool in molecular epidemiology of infectious diseases. *Annals of Epidemiology*, v. 16, n. 3, p. 157–169, 2006. ISSN 1047-2797. Disponível em: ¡⟨https://www.sciencedirect.com/science/article/pii/S1047279705001080⟩¿.

3 ALBERTS, B. et al. *Molecular Biology of the Cell*. 4th edition. ed. New York: Garland Science, 2002. ISBN 978-0815344643.

4 MORRISON, D. A. Tree Thinking: An Introduction to Phylogenetic Biology. David A. Baum and Stacey D. Smith. *Systematic Biology*, v. 62, n. 4, p. 634–637, 05 2013. ISSN 1063-5157. Disponível em: ¡⟨https://doi.org/10.1093/sysbio/syt026⟩¿.

5 FALL, A. et al. Genetic diversity and evolutionary dynamics of respiratory syncytial virus over eleven consecutive years of surveillance in Senegal. *Infection, Genetics and Evolution*, v. 91, p. 104864, jul. 2021. ISSN 15671348. Disponível em: ¡⟨https://linkinghub.elsevier.com/retrieve/pii/S1567134821001611⟩¿.

6 SHABBIR, M. Z.; RAHMAN, A.-u.; MUNIR, M. A comprehensive global perspective on phylogenomics and evolutionary dynamics of Small ruminant morbillivirus. *Scientific Reports*, v. 10, n. 1, p. 17, dez. 2020. ISSN 2045-2322. Disponível em: ¡⟨http://www.nature.com/articles/s41598-019-54714-w⟩¿.

7 HUDU, S. A. et al. Hepatitis E virus isolated from chronic hepatitis B patients in Malaysia: Sequences analysis and genetic diversity suggest zoonotic origin. *Alexandria Journal of Medicine*, v. 54, n. 4, p. 487–494, dez. 2018. ISSN 2090-5068, 2090-5076. Disponível em: ¡⟨https://www.tandfonline.com/doi/full/10.1016/j.ajme.2017.07.003⟩¿.

8 SALLARD, E. et al. Tracing the origins of SARS-COV-2 in coronavirus phylogenies: a review. *Environmental Chemistry Letters*, v. 19, n. 2, p. 769–785, abr. 2021. ISSN 1610-3653, 1610-3661. Disponível em: ¡⟨https://link.springer.com/10.1007/s10311-020-01151-1⟩¿.

9 PAEZ-ESPINO, D. et al. Diversity, evolution, and classification of virophages uncovered through global metagenomics. *Microbiome*, v. 7, n. 1, p. 157, dez. 2019. ISSN 2049-2618. Disponível em: ¡⟨https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0768-5⟩¿.

10 TANG, X. et al. Evolutionary analysis and lineage designation of SARS-CoV-2 genomes. *Science Bulletin*, v. 66, n. 22, p. 2297–2311, nov. 2021. ISSN 20959273. Disponível em: ¡⟨https://linkinghub.elsevier.com/retrieve/pii/S2095927321001250⟩¿.

11 CHO, M.; MIN, X.; SON, H. S. Analysis of evolutionary and genetic patterns in structural genes of primate lentiviruses. *Genes & Genomics*, v. 44, n. 7, p. 773–791, jul. 2022. ISSN 1976-9571, 2092-9293. Disponível em: ¡⟨https://link.springer.com/10.1007/s13258-022-01257-6⟩¿.

12 YIN, Y. et al. A systematic genotype and subgenotype re-ranking of hepatitis B virus under a novel classification standard. *Heliyon*, v. 5, n. 10, p. e02556, out. 2019. ISSN 24058440. Disponível em: ¡⟨https://linkinghub.elsevier.com/retrieve/pii/S2405844019362164⟩¿.

13 BEDOYA-PILOZO, C. H. et al. Molecular epidemiology and phylogenetic analysis of human papillomavirus infection in women with cervical lesions and cancer from the coastal region of Ecuador. *Revista Argentina de Microbiología*, v. 50, n. 2, p. 136–146, abr. 2018. ISSN 03257541. Disponível em: ¡⟨https://linkinghub.elsevier.com/retrieve/pii/S0325754117301372⟩¿.

14 POTDAR, V. et al. Phylogenetic classification of the whole-genome sequences of SARS-CoV-2 from India & evolutionary trends. *Indian Journal of Medical Research*, v. 153, n. 1, p. 166, 2021. ISSN 0971-5916. Disponível em: ¡⟨https://journals.lww.com/ijmr/Fulltext/2021/01000/Phylogenetic_classification_of_the_whole_genome.14.aspx⟩¿.

15 LICHTBLAU, D. Alignment-free genomic sequence comparison using FCGR and signal processing. *BMC Bioinformatics*, v. 20, n. 1, p. 742, dez. 2019. ISSN 1471-2105. Disponível em: ¡⟨https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3330-3⟩¿.

16 KIM, J.; CHEON, S.; AHN, I. NGS data vectorization, clustering, and finding key codons in SARS-CoV-2 variations. *BMC Bioinformatics*, v. 23, n. 1, p. 187, dez. 2022. ISSN 1471-2105. Disponível em: ¡⟨https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04718-7⟩¿.

17 DIMITROV, K. M. et al. Updated unified phylogenetic classification system and revised nomenclature for Newcastle disease virus. *Infection, Genetics and Evolution*, v. 74, p. 103917, out. 2019. ISSN 15671348. Disponível em: ¡⟨https://linkinghub.elsevier.com/retrieve/pii/S1567134819301388⟩¿.

18 PEFFERS, K. et al. A design science research methodology for information systems research. *Journal of management information systems*, Taylor & Francis, v. 24, n. 3, p. 45–77, 2007.

19 SILVA, D. D.; LOPES, E. L.; JUNIOR, S. S. B. Pesquisa Quantitativa: Elementos, Paradigmas e Definições. *Revista de Gestão e Secretariado*, v. 05, n. 01, p. 01–18, abr. 2014. ISSN 21789010, 21789010. Disponível em: ¡⟨http://www.revistagesec.org.br/ojs-2.4.5/index.php/secretariado/article/view/297⟩¿.