

Mauricio Souza Menezes

*Ferramenta Computacional para o Estudo  
da Evolução de Espécies Baseado no Uso  
de Códon*

Salvador-BA

5 de dezembro de 2022



Mauricio Souza Menezes

*Ferramenta Computacional para o Estudo  
da Evolução de Espécies Baseado no Uso  
de Códon*

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

5 de dezembro de 2022



## Folha de Aprovação

Anteprojeto sob o título provisório *Ferramenta Computacional para o Estudo da Evolução de Espécies Baseado no Uso de Códon*s 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 Marco Antônio Costa Simões professor da disciplina, em 5 de dezembro de 2022, em Salvador, Bahia.

---

Mauricio Souza Menezes  
Orientando

---

Diego Gervasio Frias Suárez

# *Sumário*

<b>1</b>	<b>Introdução</b>	p. 6
<b>2</b>	<b>Objetivos</b>	p. 8
2.1	Objetivo Geral . . . . .	p. 8
2.2	Objetivos Específicos . . . . .	p. 8
<b>3</b>	<b>Justificativas e Contribuições</b>	p. 9
<b>4</b>	<b>Metodologia</b>	p. 10
<b>5</b>	<b>Cronograma</b>	p. 11
	<b>Referências</b>	p. 13



# 1 *Introdução*

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

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.’

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 (3) 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). Nos eucariotos, a principal fonte de mutação é a radiação incidente nas gônadas, que pode ser ambiental (solo com material radioativo), artificial (radiografias e tomografias) ou radiação cósmica.

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, muitas vezes 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 (*Maximum Likelihood*) não é exclusivo da filogenia, mas sim uma abordagem estatística. A aplicação da Máxima Verossimilhança em filogenia consiste em avaliar a probabilidade de que o modelo de evolução escolhido gere os dados observados. Essa proposta foi utilizada por (4), (1), (5), (6), (7), (8), (9) e (10).

Já o

procurar pela árvore que tem a maior probabilidade de dar origem ao dado observado. No caso da filogenia os dados observados serão as características de cada organismo (a matriz de estado de características) e as hipóteses são todas as possíveis árvores. É também necessário um modelo matemático de evolução de características que será aplicado pelo ML.

Os trabalhos relacionados apresentaram vários métodos de classificação gênica. Como por exemplo (11) comparou três modelos para reconstrução de árvores filogenéticas: junção de vizinhos; máxima verossimilhança e inferência bayesiana. Para (12) e (13) foi usada a inferência bayesiana. Já a proposta do método de máxima verossimilhança foi utilizada por (4), (1), (5), (6), (7), (8), (9) e (10). Além desses, o trabalho de (14) expôs o Frequency Chaos Game Representation, (15) a floresta aleatória e (16) a junção de vizinhos.

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. Necessita-se pesquisa 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 de vírus

### **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.

### *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 serviria 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 elaborada numa pesquisa do tipo quantitativa, ou seja, com o uso de 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.(17)

## 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, 6(seis) horas diárias para realização das atividades.

	Jul	Ago	Set	Out	Nov	Dez
	•	•	•			
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 Julho até Dezembro de 2022

	Jan	Fev	Mar	Abr	Mai	Jun
Desenvolvimento da primeira versão do método	•	•	•			
Avaliação formativa do método de acordo com a metodologia proposta para o projeto			•	•		
Testar e implementar as melhorias necessárias para atingir o objetivo				•	•	
Análise e considerações do resultado final						•

Tabela 2: Cronograma de Janeiro a Junho 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: [j<https://linkinghub.elsevier.com/retrieve/pii/S1567134822000144>ç](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: [j<https://www.sciencedirect.com/science/article/pii/S1047279705001080>ç](https://www.sciencedirect.com/science/article/pii/S1047279705001080).
- 3 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: [j<https://doi.org/10.1093/sysbio/syt026>ç](https://doi.org/10.1093/sysbio/syt026).
- 4 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: [j<https://linkinghub.elsevier.com/retrieve/pii/S1567134821001611>ç](https://linkinghub.elsevier.com/retrieve/pii/S1567134821001611).
- 5 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: [j<http://www.nature.com/articles/s41598-019-54714-w>ç](http://www.nature.com/articles/s41598-019-54714-w).
- 6 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: [j<https://www.tandfonline.com/doi/full/10.1016/j.ajme.2017.07.003>ç](https://www.tandfonline.com/doi/full/10.1016/j.ajme.2017.07.003).
- 7 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: [j<https://link.springer.com/10.1007/s10311-020-01151-1>ç](https://link.springer.com/10.1007/s10311-020-01151-1).
- 8 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: [j<https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0768-5>ç](https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0768-5).
- 9 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: [j<https://linkinghub.elsevier.com/retrieve/pii/S2095927321001250>ç](https://linkinghub.elsevier.com/retrieve/pii/S2095927321001250).

- 10 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: [j<https://link.springer.com/10.1007/s13258-022-01257-6>ç](https://link.springer.com/10.1007/s13258-022-01257-6).
- 11 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: [j<https://linkinghub.elsevier.com/retrieve/pii/S1567134819301388>ç](https://linkinghub.elsevier.com/retrieve/pii/S1567134819301388).
- 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: [j<https://linkinghub.elsevier.com/retrieve/pii/S2405844019362164>ç](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: [j<https://linkinghub.elsevier.com/retrieve/pii/S0325754117301372>ç](https://linkinghub.elsevier.com/retrieve/pii/S0325754117301372).
- 14 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: [j<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3330-3>ç](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3330-3).
- 15 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: [j<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04718-7>ç](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04718-7).
- 16 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: [j<https://journals.lww.com/ijmr/Fulltext/2021/01000/Phylogenetic\\_classification\\_of\\_the\\_whole\\_genome.14.aspx>ç](https://journals.lww.com/ijmr/Fulltext/2021/01000/Phylogenetic_classification_of_the_whole_genome.14.aspx).
- 17 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: [j<http://www.revistagesec.org.br/ojs-2.4.5/index.php/secretariado/article/view/297>ç](http://www.revistagesec.org.br/ojs-2.4.5/index.php/secretariado/article/view/297).