Uma Revisão de Literatura sobre *Ferramentas Computacionais para o Estudo da Evolução de Espécies Baseado no Uso de Códons*

Mauricio Souza Menezes Teste

[1]Departamento de Ciências Exatas e da Terra, Campus I
Universidade do Estado da Bahia (UNEB)
Salvador, Bahia, Brasil.

mauriciosm95@gmail.com

Resumo. *Esta Revisão Sistemática (RS) dos métodos de classificação genética de sequências implementados e aplicados até o momento. Vários métodos diferentes foram identificados, sugerindo que nenhum deles foi considerado indubitável.*

Abstract. *This Systematic Review (SR) of sequence genetic classification methods implemented and applied to date. Several different methods were identified, suggesting that none of them was considered indubitable.*

1. Introdução

As análises de reconstrução filogenética é uma ferramenta muito utilizada no campo das Ciências Biológicas. O seu uso busca identificar e compreender as relações evolutivas entre diferentes formas de vida no planeta, pois ela busca compreender a evolução de agentes causadores de doenças. Segundo [Behl et al. 2022] 'existe uma tendência histórica, onde os vírus e bactérias sofrem mutações no decorrer do tempo, encontrando mecanismos para infectar as células humanas.' Este artigo apresenta uma RS da literatura referente aos métodos computacionais utilizados para realizar o estudo de especiação genômica. Os artigos foram obtidos através da busca realizada em quatro bases de dados (PubMed, ScienceDirect, Scopus e Springer Link) levando em conta a principal questão: Quais métodos são utilizados para realizar o estudo de especiação? Foram incluídos todos os artigos que atenderam aos critérios definidos.

2. Relato da Revisão de Literatura

Logo após definir os objetivos da pesquisa, foi primordial conhecer os pormenores que estavam relacionados aos mesmos, e para isso foi necessário realizar uma RS. Dessa forma, foi utilizado um protocolo que serviria de base no processo para a obtenção dos estudos que deveriam ser analisados, os quais foram definidos e analisados através de: Objetivos, questões, palavras-chave, strings de busca, fontes e critérios para inclusão e exclusão.

2.1. Objetivo da Pesquisa

A pesquisa teve como objetivo verificar e conhecer a existência de métodos utilizados para a analise da evolução de espécies com base no uso de códons, a identificação do que já foi feito para sanar o problema identificado e também se era viável ou não o desenvolvimento de uma nova solução para classificação gênica.

2.2. Questões de Pesquisa

Foram definidas duas questões de pesquisa. A primeira e principal questionava quais métodos são utilizados para realizar o estudo o estudo de especiação. Enquanto a segunda era se alguma dessas metodologias é baseada no uso/frequência de códons. As questões de pesquisa foram escolhidas com base naquilo que se queria entender ao final da RS.

2.3. Repositório de Busca de Dados

As bases precisavam conter trabalhos relevantes relacionados a bioinformática, completos e gratuitos. No entanto, em alguns casos, esse acesso se deu através por meio do portal da Comunidade Acadêmica Federada (CAFe). Deste modo, as seguintes bases de dados acadêmicos foram selecionadas:

- PubMed
- ScienceDirect
- Scopus
- Springer Link

Para o refinamento das consultas foram utilizados os filtros que serão apresentados a seguir:

1. ScienceDirect
   (a) Article type: Review Articles e Research Articles
   (b) Access type: Open access e Open Archive
   (c) Years: 2018 até 2022
2. Springer Link
   (a) Include preview-only content: Desmarcado
   (b) Content type: Article
   (c) Language: English
   (d) Years: 2018 até 2022
3. Scopus
   (a) Open access: All open access
   (b) Years: 2018 até 2022

2.4. Palavras-chave e Strings de Busca

As palavras-chave, com seus sinônimos e correlatos em inglês, selecionadas para a busca nas bases de dados, com o objetivo de encontrar o máximo de estudos relevantes, foram as seguintes:

- Bioinformatics e Bioinformática
- Codon e códon
- Gene
- Phylogeny e filogenia
- Classification, classificação, genotyping, genotipagem, subtyping, subtipagem, typing e tipagem
- Viral

Com base nas palavras-chave, foi criada a string de busca a seguir, que foi utilizada em todas as bases de dados.

- ((‘bioinformatics’) AND (‘codon’) AND (‘phylogeny’) AND (‘typing’ OR ‘classification’ OR ‘genotyping’ OR ‘subtyping’) AND (‘gene’) AND (‘viral’))

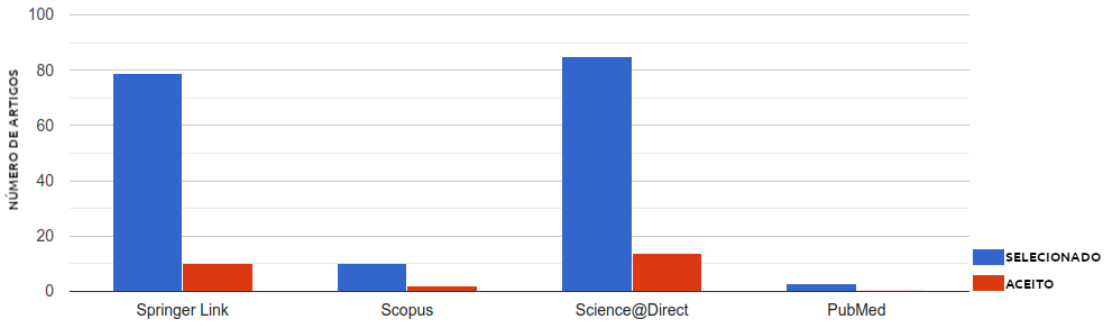

Figura 1. Artigos Selecionados e Aceitos por Base.

2.5. Critérios de seleção

Para realizar o processo de seleção dos estudos foram definidos critérios, podendo ser de inclusão ou de exclusão. Os critérios de inclusão foram os seguintes:

- I1: O estudo contém as palavras-chave definidas na pesquisa no resumo, título ou palavras-chave.
- I2: Aborda algum método de especiação.

Os critérios de exclusão foram os seguintes:

- E1: O tema do estudo não está relacionado à especiação.
- E2: O tema do estudo não é pertinente com a área/objetivos da pesquisa.
- E3: O estudo está duplicado.

3. Resultados Parciais

3.1. Análise Quantitativa dos Resultados

A aplicação das strings nas bases de dados selecionadas, resultou em um total geral de 177 (cento e setenta e sete) artigos, os quais foram oriundos da fonte PubMed 3 (três) artigos, da fonte ScienceDirect 85 (oitenta e cinco) artigos, da fonte Scopus 10 (dez) artigos e da fonte Springer Link 79 (setenta e nove) artigos.
Após o processo de obtenção, foi iniciada a seleção dos artigos, onde foi realizada a leitura do título, resumo e palavras-chave, aplicando os critérios de inclusão e exclusão. Esse processo resultou em: 26 (vinte e seis) artigos aceitos; 147 (cento e quarenta e sete) artigos rejeitados; e 4 (quatro) artigos duplicados. Os resultados são apresentados nos gráficos nas figuras 1 e 2.

3.2. Análise Qualitativa dos Resultados

Na etapa de extração, realizada com o auxílio da ferramenta Perform Systematic Literature Reviews (Parsifal), houve o levantamento de informações para ajudar a responder as questões de pesquisa, 4(quatro) para a principal e 1(uma) para a secundaria, apresentadas a seguir:

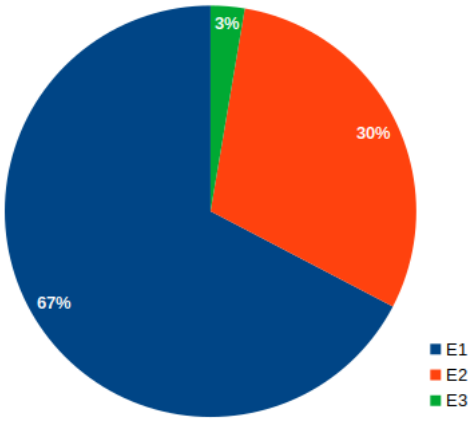

Figura 2. Artigos Excluídos por Critério.

- P1: Qual método de classificação genética foi utilizado?
- P2: O método necessitava de algum treinamento?
- P3: Se necessita de algum treinamento, o mesmo era supervisionado?
- P4: O método necessitava de uma árvore de referência supervisionada?
- S1: Alguma dessas metodologias é baseada no uso/frequência de códons?

Essas informações foram importantes também para a construção da planilha-resumo de resultados. Também é importante salientar, que após a leitura completa, 12(doze) dos trabalhos se enquadraram em um dos critérios de exclusão apresentados anteriormente.

Se tratando dos métodos de classificação gênica utilizados, o trabalho [Dimitrov et al. 2019] apresentou e comparou três modelos para reconstrução de árvores filogenéticas: junção de vizinhos; máxima verossimilhança e inferência bayesiana. Para [Yin et al. 2019] e [Bedoya-Pilozo et al. 2018] foi usada a inferência bayesiana. Temos que em [Fall et al. 2021], [Behl et al. 2022], [Shabbir et al. 2020], [Hudu et al. 2018], [Sallard et al. 2021], [Pae e [Cho et al. 2022] foi aplicada a máxima verossimilhança. Os demais estudos apresentaram metodologias distintas: No trabalho de [Lichtblau 2019] foi Frequency Chaos Game Representation (FCGR); [Kim et al. 2022] a floresta aleatória e [Potdar et al. 2021] a junção de vizinhos. Além disso, levando em conta a necessidade de treinamento, [Lichtblau 2019] utilizou redes neurais, onde não houve treinamento supervisionado ou necessidade de uma árvore de referência supervisionada.

Já em relação metodologias baseadas em uso/frequência de códons, nenhum dos trabalhos apresentou tal proposta. O que mais se aproximou disso foi [Cho et al. 2022] que segundo ele 'realizou um agrupamento da análise do uso de códons com base nos resultados das árvores filogenéticas'.

4. Conclusões

A possibilidade do desenvolvimento de uma ferramenta de classificação gênica com base no uso/frequência de códons se mostrou possível, tendo em vista que os métodos encontrados não realizavam tal metodologia. Sendo assim, uma nova sistematização para

classificação de sequências genéticas poderá contribuir com avanços na otimização de classificação de grandes conjuntos de dados para determinar a sua caracterização genômica.

Referências

Ahmad, S. U., Hafeez Kiani, B., Abrar, M., Jan, Z., Zafar, I., Ali, Y., Alanazi, A. M., Malik, A., Rather, M. A., Ahmad, A., and Khan, A. A. (2022). A comprehensive genomic study, mutation screening, phylogenetic and statistical analysis of SARS-CoV-2 and its variant omicron among different countries. *Journal of Infection and Public Health*, 15(8):878–891.

Bedoya-Pilozo, C. H., Medina Magües, L. G., Espinosa-García, M., Sánchez, M., Parrales Valdiviezo, J. V., Molina, D., Ibarra, M. A., Quimis-Ponce, M., España, K., Párraga Macias, K. E., Cajas Flores, N. V., Orlando, S. A., Robalino Penaherrera, J. A., Chedraui, P., Escobar, S., Loja Chango, R. D., Ramirez-Morán, C., Espinoza-Caicedo, J., Sánchez-Giler, S., Limia, C. M., Alemán, Y., Soto, Y., Kouri, V., Culasso, A. C., and Badano, I. (2018). 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*, 50(2):136–146.

Behl, A., Nair, A., Mohagaonkar, S., Yadav, P., Gambhir, K., Tyagi, N., Sharma, R. K., Butola, B. S., and Sharma, N. (2022). Threat, challenges, and preparedness for future pandemics: A descriptive review of phylogenetic analysis based predictions. *Infection, Genetics and Evolution*, 98:105217.

Cho, M., Min, X., and Son, H. S. (2022). Analysis of evolutionary and genetic patterns in structural genes of primate lentiviruses. *Genes & Genomics*, 44(7):773–791.

Dimitrov, K. M., Abolnik, C., Afonso, C. L., Albina, E., Bahl, J., Berg, M., Briand, F.-X., Brown, I. H., Choi, K.-S., Chvala, I., Diel, D. G., Durr, P. A., Ferreira, H. L., Fusaro, A., Gil, P., Goujgoulova, G. V., Grund, C., Hicks, J. T., Joannis, T. M., Torchetti, M. K., Kolosov, S., Lambrecht, B., Lewis, N. S., Liu, H., Liu, H., McCullough, S., Miller, P. J., Monne, I., Muller, C. P., Munir, M., Reischak, D., Sabra, M., Samal, S. K., Servan de Almeida, R., Shittu, I., Snoeck, C. J., Suarez, D. L., Van Borm, S., Wang, Z., and Wong, F. Y. (2019). Updated unified phylogenetic classification system and revised nomenclature for Newcastle disease virus. *Infection, Genetics and Evolution*, 74:103917.

Fall, A., Elawar, F., Hodcroft, E. B., Jallow, M. M., Toure, C. T., Barry, M. A., Kiori, D. E., Sy, S., Diaw, Y., Goudiaby, D., Niang, M. N., and Dia, N. (2021). Genetic diversity and evolutionary dynamics of respiratory syncytial virus over eleven consecutive years of surveillance in Senegal. *Infection, Genetics and Evolution*, 91:104864.

Hudu, S. A., Niazlin, M. T., Nordin, S. A., Harmal, N. S., Tan, S. S., Omar, H., Shahar, H., Mutalib, N. A., and Sekawi, Z. (2018). Hepatitis E virus isolated from chronic hepatitis B patients in Malaysia: Sequences analysis and genetic diversity suggest zoonotic origin. *Alexandria Journal of Medicine*, 54(4):487–494.

Kim, J., Cheon, S., and Ahn, I. (2022). NGS data vectorization, clustering, and finding key codons in SARS-CoV-2 variations. *BMC Bioinformatics*, 23(1):187.

Lichtblau, D. (2019). Alignment-free genomic sequence comparison using FCGR and signal processing. *BMC Bioinformatics*, 20(1):742.

Paez-Espino, D., Zhou, J., Roux, S., Nayfach, S., Pavlopoulos, G. A., Schulz, F., McMahon, K. D., Walsh, D., Woyke, T., Ivanova, N. N., Eloe-Fadrosh, E. A., Tringe, S. G., and Kyrpides, N. C. (2019). Diversity, evolution, and classification of virophages uncovered through global metagenomics. *Microbiome*, 7(1):157.

Potdar, V., Vipat, V., Ramdasi, A., Jadhav, S., Pawar-Patil, J., Walimbe, A., Patil, S., Choudhury, M., Shastri, J., Agrawal, S., Pawar, S., Lole, K., Abraham, P., and Cherian, S. (2021). Phylogenetic classification of the whole-genome sequences of SARS-CoV-2 from India & evolutionary trends. *Indian Journal of Medical Research*, 153(1):166.

Sallard, E., Halloy, J., Casane, D., Decroly, E., and van Helden, J. (2021). Tracing the origins of SARS-COV-2 in coronavirus phylogenies: a review. *Environmental Chemistry Letters*, 19(2):769–785.

Shabbir, M. Z., Rahman, A.-u., and Munir, M. (2020). A comprehensive global perspective on phylogenomics and evolutionary dynamics of Small ruminant morbillivirus. *Scientific Reports*, 10(1):17.

Tang, X., Ying, R., Yao, X., Li, G., Wu, C., Tang, Y., Li, Z., Kuang, B., Wu, F., Chi, C., Du, X., Qin, Y., Gao, S., Hu, S., Ma, J., Liu, T., Pang, X., Wang, J., Zhao, G., Tan, W., Zhang, Y., Lu, X., and Lu, J. (2021). Evolutionary analysis and lineage designation of SARS-CoV-2 genomes. *Science Bulletin*, 66(22):2297–2311.

Yin, Y., He, K., Wu, B., Xu, M., Du, L., Liu, W., Liao, P., Liu, Y., and He, M. (2019). A systematic genotype and subgenotype re-ranking of hepatitis B virus under a novel classification standard. *Heliyon*, 5(10):e02556.

5. Planilha-resumo de Resultados

Na 1 é apresentada a planilha-resumo de resultados com os trabalhos aceitos no processo de seleção. Assim, para cada artigo foi extraído a sua identificação, os critérios de inclusão ou exclusão que foram aplicados, uma breve descrição e as respostas para as perguntas apresentadas na seção de análise qualitativa dos resultados. Alguns campos foram respondidos com S (sim), N (não) e O (Omitido).

Tabela 1. Planilha-resumo dos trabalhos selecionados.

Identificação do Trabalho	I1	I2	E1	E2	E3	Descrição	P1	P2	P3	P4	S1
Ahmad, S. U., Hafeez Kiani, B., Abrar, M., Jan, Z., Zafar, I., Ali, Y., Alanazi, A. M., Malik, A., Rather, M. A., Ahmad, A., and Khan, A. A. (2022). A comprehensive genomic study, mutation screening, phylogenetic and statistical analysis of SARS-CoV-2 and its variant omicron among different countries. *Journal of Infection and Public Health*, 15(8):878–891		X				O estudo tem como objetivo investigar e analisar a mutação d 157 genomas de SARS-Cov-2 e suas variantes Delta e Omicron.	Neighbor-Joining	N	N	N	N
Yin, Y., He, K., Wu, B., Xu, M., Du, L., Liu, W., Liao, P., Liu, Y., and He, M. (2019). A systematic genotype and subgenotype re-ranking of hepatitis B virus under a novel classification standard. *Heliyon*, 5(10):e02556		X				O estudo apresentou a reconstrução a filogenia do HBV com base em 4.429 sequências completas.	Inferência bayesiana	N	N	N	N
Lichtblau, D. (2019). Alignment-free genomic sequence comparison using FCGR and signal processing. *BMC Bioinformatics*, 20(1):742		X				O estudo apresentou a utilização de métodos livres de alinhamento de comparação genômica para escalonamento de grandes conjuntos de dados de sequências de nucleotídeos.	Frequency Chaos Game Representation (FCGR).	S	N	O	N

Continua na próxima página

Tabela 1 – continuação da página anterior

Identificação do Trabalho	I1	I2	E1	E2	E3	Descrição	P1	P2	P3	P4	S1
Cho, M., Min, X., and Son, H. S. (2022). Analysis of evolutionary and genetic patterns in structural genes of primate lentiviruses. *Genes & Genomics*, 44(7):773–791	X	X				O estudo teve como objetivo confirmar o padrão geral de uso de códons e explorar as características evolutivas e genéticas comumente ou especificamente expressas em HIV1, HIV2 e SIV.	Máxima verossimilhança	N	O	O	S
Paez-Espino, D., Zhou, J., Roux, S., Nayfach, S., Pavlopoulos, G. A., Schulz, F., McMahon, K. D., Walsh, D., Woyke, T., Ivanova, N. N., Eloe-Fadrosh, E. A., Tringe, S. G., and Kyrpides, N. C. (2019). Diversity, evolution, and classification of virophages uncovered through global metagenomics. *Microbiome*, 7(1):157		X				O trabalho examinou uma coleção com 14.000 metagenomas, identificando 44.221 sequências de virófagos, das quais 328 era genomas completos ou quase completos. Nesses foi realizada a analise genômica comparativa.	Alinhamento concatenado.	N	N	N	N
Tang, X., Ying, R., Yao, X., Li, G., Wu, C., Tang, Y., Li, Z., Kuang, B., Wu, F., Chi, C., Du, X., Qin, Y., Gao, S., Hu, S., Ma, J., Liu, T., Pang, X., Wang, J., Zhao, G., Tan, W., Zhang, Y., Lu, X., and Lu, J. (2021). Evolutionary analysis and lineage designation of SARS-CoV-2 genomes. *Science Bulletin*, 66(22):2297–2311		X				O estudo analisou variantes de nucleotídeo único (SNVs) em 121.618 genomas de SARS-CoV-2 de alta qualidade.	Máxima parcimônia	N	N	N	N

Continua na próxima página

Tabela 1 – continuação da página anterior

Identificação do Trabalho	I1	I2	E1	E2	E3	Descrição	P1	P2	P3	P4	S1
Fall, A., Elawar, F., Hodcroft, E. B., Jallow, M. M., Toure, C. T., Barry, M. A., Kiori, D. E., Sy, S., Diaw, Y., Goudiaby, D., Niang, M. N., and Dia, N. (2021). Genetic diversity and evolutionary dynamics of respiratory syncytial virus over eleven consecutive years of surveillance in Senegal. *Infection, Genetics and Evolution*, 91:104864		X				O estudo analisou a diversidade genética e a dinâmica evolutiva do HSRV no Senegal com dados coletados entre 2008 e 2018 com o objetivo de compreender a base da evolução molecular das cepas.	Máxima verossimilhança.	N	N	N	N
Hudu, S. A., Niazlin, M. T., Nordin, S. A., Harmal, N. S., Tan, S. S., Omar, H., Shahar, H., Mutalib, N. A., and Sekawi, Z. (2018). Hepatitis E virus isolated from chronic hepatitis B patients in Malaysia: Sequences analysis and genetic diversity suggest zoonotic origin. *Alexandria Journal of Medicine*, 54(4):487–494		X				O estudo caracterizou a análise genômica comparativa do HEV de 82 pacientes nos anos de 2015 e 2016.	Máxima verossimilhança.	N	N	N	N

Continua na próxima página

Tabela 1 – continuação da página anterior

Identificação do Trabalho	I1	I2	E1	E2	E3	Descrição	P1	P2	P3	P4	S1
Bedoya-Pilozo, C. H., Medina Magües, L. G., Espinosa-García, M., Sánchez, M., Parrales Valdiviezo, J. V., Molina, D., Ibarra, M. A., Quimis-Ponce, M., España, K., Párraga Macias, K. E., Cajas Flores, N. V., Orlando, S. A., Robalino Penaherrera, J. A., Chedraui, P., Escobar, S., Loja Chango, R. D., Ramirez-Morán, C., Espinoza-Caicedo, J., Sánchez-Giler, S., Limia, C. M., Alemán, Y., Soto, Y., Kouri, V., Culasso, A. C., and Badano, I. (2018). 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*, 50(2):136–146		X				O estudo analisou a evolução das variantes do HPV mais prevalentes com base em 166 amostras caracterizando-as através da filogenia e coalescência.	Inferência bayesiana	N	N	N	N
Kim, J., Cheon, S., and Ahn, I. (2022). NGS data vectorization, clustering, and finding key codons in SARS-CoV-2 variations. *BMC Bioinformatics*, 23(1):187	X	X				O estudo propôs métodos para vetorizar os dados da sequência, realizar análises de agrupamento e visualizar os resultados com métodos de aprendizagem de máquina.	Floresta aleatória	S	S	O	N

Continua na próxima página

Tabela 1 – continuação da página anterior

Identificação do Trabalho	I1	I2	E1	E2	E3	Descrição	P1	P2	P3	P4	S1
Potdar, V., Vipat, V., Ramdasi, A., Jadhav, S., Pawar-Patil, J., Walimbe, A., Patil, S., Choudhury, M., Shastri, J., Agrawal, S., Pawar, S., Lole, K., Abraham, P., and Cherian, S. (2021). Phylogenetic classification of the whole-genome sequences of SARS-CoV-2 from India & evolutionary trends. *Indian Journal of Medical Research*, 153(1):166		X				O estudo fornece uma integração das classificações filogenéticas existentes, e descreve as tendências evolutivas das cepas de SARS-CoV-2 que circulam na Índia. Foi realizada a análise de 3.014 sequências indianas de SARS-CoV-2.	Junção de vizinhos	N	N	N	N
Behl, A., Nair, A., Mohagaonkar, S., Yadav, P., Gambhir, K., Tyagi, N., Sharma, R. K., Butola, B. S., and Sharma, N. (2022). Threat, challenges, and preparedness for future pandemics: A descriptive review of phylogenetic analysis based predictions. *Infection, Genetics and Evolution*, 98:105217		X				O estudo apresenta uma análise evolutiva da doença infecciosa através de análises filogenéticas.	Máxima verossimilhança.	N	N	N	N
Sallard, E., Halloy, J., Casane, D., Decroly, E., and van Helden, J. (2021). Tracing the origins of SARS-COV-2 in coronavirus phylogenies: a review. *Environmental Chemistry Letters*, 19(2):769–785		X				O estudo apresentou uma discussão bre a origem, natural ou sintética, do SARS-CoV-2, com base em inferências filogenéticas, análises de sequências e relações estrutura-função das proteínas do coronavírus.	Inferências filogenéticas	N	N	N	N

Continua na próxima página

Tabela 1 – continuação da página anterior

Identificação do Trabalho	I1	I2	E1	E2	E3	Descrição	P1	P2	P3	P4	S1
Dimitrov, K. M., Abolnik, C., Afonso, C. L., Albina, E., Bahl, J., Berg, M., Briand, F.-X., Brown, I. H., Choi, K.-S., Chvala, I., Diel, D. G., Durr, P. A., Ferreira, H. L., Fusaro, A., Gil, P., Goujgoulova, G. V., Grund, C., Hicks, J. T., Joannis, T. M., Torchetti, M. K., Kolosov, S., Lambrecht, B., Lewis, N. S., Liu, H., Liu, H., McCullough, S., Miller, P. J., Monne, I., Muller, C. P., Munir, M., Reischak, D., Sabra, M., Samal, S. K., Servan de Almeida, R., Shittu, I., Snoeck, C. J., Suarez, D. L., Van Borm, S., Wang, Z., and Wong, F. Y. (2019). Updated unified phylogenetic classification system and revised nomenclature for Newcastle disease virus. *Infection, Genetics and Evolution*, 74:103917		X				O estudo propôs um sistema de classificação para facilitar a nomenclatura em estudos da evolução e epidemiologia de vírus da doença de Newcastle.	Junção de vizinhos, máxima verossimilhança e inferência bayesiana.	N	N	N	N