Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 12, n. 1, 2026.

Trabalho apresentado no XLIV CNMAC, Fundação Getulio Vargas - Rio de Janeiro - RJ, 2025.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Método de Clusterização hibrida Baseado em Logica Fuzzy para Espécies de Kinetoplastea Usando Sequências Não Alinhadas

Antonio R. A. Pereira[1], Flávio L. de Melo[2], Marcello G.Teixeira[3]
UFRJ, Rio de Janeiro, RJ
Maria A. Dario[4], Márcio G. Pavan[5], Ana M. Jansen [6], Samanta C. C. Xavier[7]
IOC/FIOCRUZ, Rio de Janeiro, Rj

Resumo. Este artigo propõe uma metodologia híbrida de clusterização, combinando algoritmos *fuzzy c-means* e hierárquico, para análise taxonômica de espécies da classe Kinetoplastea, utilizando sequências de DNA não alinhadas. O pré-processamento envolveu cálculo de distâncias baseado em *k-mers*, gerando uma matriz de distâncias. Os resultados identificaram cinco grupos geneticamente distintos, validados pelos métodos de correlação cophenetica, coeficiente de partição fuzzy e inercia, refletindo relações evolutivas e distinções entre parasitas e espécies de vida livre. O método apresentou resultados promissores, oferecendo uma alternativa para estudos taxonômicos em biologia computacional.

Palavras-chave. Aprendizado de Maquina, Lógica Fuzzy, Clusterização, Taxonomia, Sequências não Alinhadas

1 Introdução

A cada dia são observados avanços em inteligencia artificial (AI) aplicados em muitas áreas, seja na aplicação acadêmica ou no dia a dia. Aprendizado de máquina (AM) é uma área da IA que tem como objetivo desenvolver técnicas computacionais sobre aprendizado, capazes de adquirir conhecimento de forma automática [10]. Na área da saúde, AM já foi utilizada para predição de genes marcadores de doenças [7], para análise de desempenho de sistemas de vigilância da febre maculosa no Rio de Janeiro [8] e diagnóstico precoce por imagem de doenças mamárias [13]. Técnicas de AM também já foram utilizadas para diferenciação de arbovírus em mosquitos *Aedes aegypti* a partir de espectros de infra-vermelho próximo [6].

Em AM, a taxonomia é considerado um problema de clusterização no qual o objetivo é agrupar objetos mais similares entre si, e separar os objetivos mais distintos [12]. Eles têm sido empregados na taxonomia [4], [9], utilizando mapas auto-organizados hiperbólicos (hyperbolic Self-Organizing maps (H^2 SOMs)) como ferramenta para agrupar pequenos fragmentos de DNA de 350 organismos procariotos em seis níveis taxonômicos. [2] utilizou AM para identificar linhagens distintas de *Mycobacterium tuberculosis*, fornecendo uma ferramenta on-line útil na classificação de novas

[1]antonio.revail@ppgi.ufrj.br
[2]fmello@poli.ufrj.br
[3]marcellogt@dcc.ufrj.br
[4]maria21dario@gmail.com
[5]marciopavan@gmail.com
[6]anamariajansen2@gmail.com
[7]sam.azeredo@gmail.com

espécies. Já [5] utilizou algoritmos de AM não supervisionados, para delimitação de uma espécie de aracnídeo com alta estrutura genética populacional.

A clusterização baseada em sequências livres de alinhamento tem se tornado uma abordagem cada vez mais relevante, especialmente devido às limitações das análises tradicionais baseadas em alinhamentos. Segundo [15], essas abordagens enfrentam desafios como o alto custo computacional, a natureza NP-difícil do alinhamento múltiplo e a dependência de parâmetros, principalmente na etapa de inicialização. Para superar essas limitações, este trabalho propoe o uso de algoritmos de clusterização baseados em lógica fuzzy [3], que oferecem maior flexibilidade e são mais adequados para lidar com a incerteza e a imprecisão presentes nos dados biológicos. Além disso, empregou-se a clusterização hierárquica [11] para inferir relações de proximidade entre espécies. Diferentemente das abordagens convencionais, nossa metodologia trabalha diretamente com sequências de nucleotídeos não alinhadas de espécies da classe Kinetoplastea, permitindo uma análise mais robusta e adaptável à complexidade evolutiva desse grupo.

As espécies de Kinetoplastea analisadas nesse trabalho incluem: *Trypanosoma cruzi* (I, TcI, TcIa, II, IV e Tcbat), *Marinkellei* , *Dionisii*, *Rangeli*, *Trypanosomasp,Janseni*, *Neobat*, *Lewisi*, *Cascavelli*, *Minasense*, *Crithidia*, *Leishmania_ infantum*, *Neobodo_ Designis*, *Parabodo_ Caudatus*, *Parabodo* e *Bodonidae*.

2 Metodologia

2.1 Pré Processamento

Sequências sem alinhamento em formato *.fasta* foram pré processadas para obter uma matriz de distância. O método baseado em frequência de palavras entre duas sequência de nucleotídeos X e Y de tamanhos diferentes, parte do princípio que sequências semelhantes compartilham palavras semelhantes [15]. Os $k-mers$, são subsequências de comprimento k onde $k \in \mathbb{N}$. Esse processo é dividido em três etapas.

Primeiro, as sequências comparadas foram divididas em coleções de palavras únicas de um determinado comprimento. Como exemplo vamos considerar duas sequencias de DNA:

$$X = ATGTGTG$$
$$Y = CATGTG$$

Separando-as em sequências de três nucleotídeos $(3-mers)$. Assim foram produzidos dois conjuntos de palavras:

$$W_X = \{ATG, TGT, GTG, TGT, GTC\}$$
$$W_Y = \{CAT, ATG, TGT, GTG\}$$

Fazendo a união dos elementos dos conjuntos W_X e W_Y, excluindo as palavras repetidas, tem-se:

$$W_3 = W_X^3 \cup W_Y^3 = \{CAT, ATG, TGT, GTG\} \tag{1}$$

O segundo passo consistiu em construir dois vetores com dimensão igual a de W_3, contando o número de vezes que cada palavra de W_3 aparece em W_X e W_Y. Com isso determinou-se dois vetores:

$$C_3^X = (0, 1, 2, 2)$$
$$C_3^Y = (1, 1, 1, 1)$$

DOI: 10.5540/03.2026.012.01.0317

Finalmente a terceira etapa consistiu em calcular a dissimilaridade ente C_3^X e C_3^Y.

$$D_3^{X,Y} = ||C_3^X - C_3^Y|| = \sqrt{(0-1)^2 + (1-1)^2 + (2-1)^2 + +(2-1)^2} = \sqrt{3} = 1,73 \qquad (2)$$

Diversas medidas de distância podem ser aplicadas, porém, neste trabalho, foi utilizado a distância euclidiana.

A matriz de distância, com todos os valores, foi desenvolvida em `Python` usando as bibliotecas `csv`, `math`, `Bio.SeqIO` e `numpy`. Essa matriz serve como entrada no algoritmos que sera exposto na subseção seguinte.

2.2 Clusterização

O código para clusterização foi desenvolvido em `Python` usando as bibliotecas `numpy`, `pandas`, `skfuzzy`, `scipy`, `matplotlib` e `sklearn`. As configurações iniciais incluíram o arquivo de entrada, onde a matriz de distâncias está armazenada, e os parâmetros para o número de clusters e iterações a serem realizados.

A matriz de distâncias foi carregada a partir de um arquivo *.csv* e normalizada usando o método *MinMaxScaler*. Essa etapa garante que os valores estejam em uma escala uniforme, essencial para o desempenho do algoritmo de *clustering*.

O código é uma mescla de dois algoritmos o *fuzzy c-means* e o hierárquico, visando obter o melhor dos dois algoritmos para melhor análise dos agrupamentos. A análise hierárquica é feita por meio da construção de uma matriz de dissimilaridade derivada dos graus de pertinência. Essa matriz é usada para calcular uma matriz de ligação hierárquica, que serve de base para a criação do dendrograma.

Foram calculadas duas inércias complementares para avaliar o clustering: as inércias fuzzy (WCSS - Within-Cluster Sum of Squares) e a do dendrograma (BSS - Between-Cluster Sum of Squares) [1]. A métrica (BSS) foi calculada como a soma das distâncias quadráticas entre os centróides de cada cluster e o centróide global (média geral dos dados). A inércia fuzzy (WCSS) é a soma ponderada das distâncias quadráticas entre os dados e os centróides dos clusters, considerando os graus de pertinência. Juntas, essas métricas equilibram coesão interna e distinção entre clusters.

A cada uma das dez iterações, foram geradas e salvos gráficos do dendrograma e dos graus de pertinência dos elementos, organizados em um diretório específico. Os resultados de cada iteração incluíram métricas como: o coeficiente cophenético, o coeficiente de partição fuzzy (FPC) e inércia total (sendo a soma $WCSS + BSS$). Essas foram armazenadas em um arquivo *.csv*.

Ao final do processo, um gráfico adicional foi criado para mostrar a variação da inércia total ao longo das iterações, evidenciando a evolução do processo de clustering. Todos os resultados e gráficos foram salvos em arquivos organizados, e o código fornece um resumo dos locais onde esses dados podem ser encontrados.

3 Resultados

A análise dos clusters foi conduzida utilizando as métricas de validação mencionadas na seção anterior, complementadas pela avaliação de especialistas do Laboratório de Biologia de Tripanosomatídeos (IOC-FIOCRUZ).

Dentre os resultados obtidos, aquele que melhor agrupou as espécies está apresentado na Figura 1. A correlação cophenética desse agrupamento é de 0.9613, indicando uma forte correspondência entre o dendrograma e as distâncias reais entre os pontos de dados. A correlação cophenética aplicada ao dendrograma, varia de 0 a 1. Valores próximos de 1 indicam uma clusterização bem ajustada [14] por isso reforça a alta qualidade do agrupamento. Para a parte relacionada aos graus

de pertinência foi usado o FPC , que uma medida de validação que assume valores no intervalo $[0, 1]$, sendo 1 o melhor apresenta 0.6787 para o FPC.

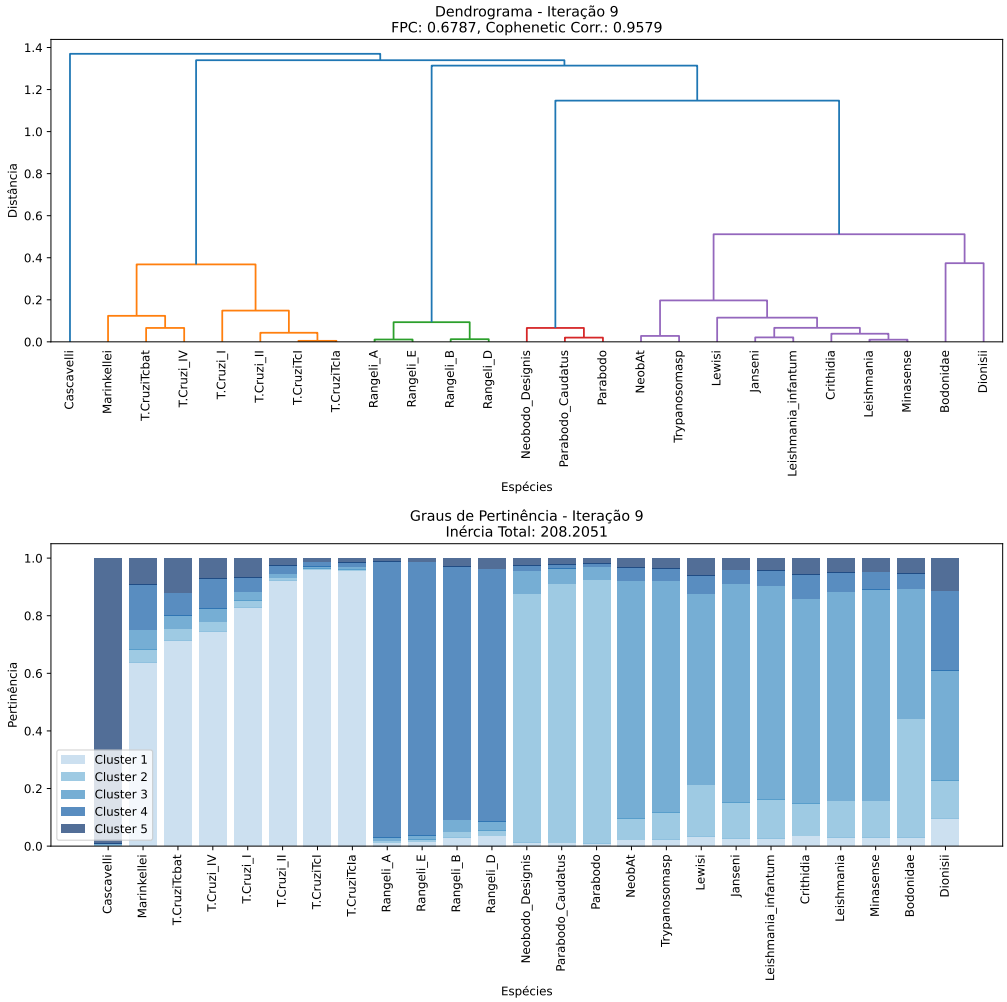

Figura 1: Dendrograma e distribuição fuzzy referente a clusterização taxonômica das espécies da classe Kinetoplastea. Fonte: O autor.

O gráfico da Figura 2 que mostra a inércia total relativa as iterações mede o quão compactos os clusters estão, ou seja, a soma das distâncias quadradas entre os pontos e seus respectivos centróides. Quanto menor a inércia, mais próximos os pontos estão do centro do cluster. A inércia total indica qual a melhor distribuição das espécies em cada cluster. Observando os valores, a inércia total oscila levemente, mantendo-se entre 208.190 e 208.215. Isso sugere que o algoritmo de clusterização estabilizou rapidamente e não há mudanças significativas na qualidade da partição após algumas iterações. Esse comportamento pode indicar que os centróides já convergiram para uma solução próxima do ótimo. Sendo assim, depois dessas análises, e da discussão com especialistas sugerimos que o melhor agrupamento é o que aparece na iteração 9, apresentada na Figura 2. A escolha por esse agrupamento deve-se à sua capacidade de separar espécies de vida livre dos parasitas e, entre os parasitas, distinguir *T. cruzi*, *Rangeli* e demais espécies, refletindo melhor

DOI: 10.5540/03.2026.012.01.0317

suas relações biológicas.

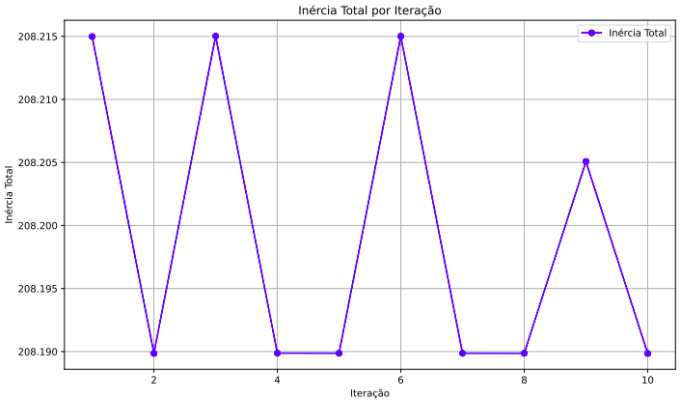

Figura 2: Gráfico de inércia total por interação. Fonte: O autor.

Os clusters apresentados na Figura 1 refletidos na Tabela 1, foram gerados automaticamente pelo algoritmo de agrupamento, sem um critério biológico pré-definido. No entanto, ao analisar os grupos formados, pode-se identificar padrões interessantes que refletem proximidade genética entre as espécies.

O **Cluster 1** é composto apenas por *Cascavelli*, que infecta serpentes. Essa espécie está separada das demais por uma distância de aproximadamente 1.2, indicando que é geneticamente distinta dos outros tripanosomas.

O **Cluster 2** reúne diferentes sequências de genótipos e subespécie de *Trypanosoma cruzi*, incluindo *T.cruzi*_II, *T.cruzi*TCI, *T.cruzi*TCIa, entre outras. Essas espécies estão conectadas por distâncias pequenas, entre 0.2 e 0.4, o que indica alta similaridade genética. Esse agrupamento é esperado, pois *T. cruzi* compartilha um ancestral comum bem definido, caracterizando um grupo monofilético.

O **Cluster 3** é formado por subgrupos de *Trypanosoma rangeli* (*Rangeli* (A, B, D e E). As distâncias intra cluster variaram entre em valores inferires a 0.2, sugerindo que essas variantes pertencem a uma espécie, mas com diferenças genéticas consideráveis. Esse grupo está mais distante de *T. cruzi*, reforçando que são espécies distintas, embora compartilhem um ancestral comum mais remoto.

O **Cluster 4** inclui espécies de vida livre, como *Neobodo_ Designis* e *Parabodo_ caudatus*. Essas espécies apresentam distâncias superiores a 1.0 em relação aos outros clusters, refletindo uma grande divergência genética, o que sugere que estão taxonomicamente distantes dos tripanosomas parasitas.

Por fim, o **Cluster 5** agrupa uma miscelânea de espécies parasitas, incluindo *Neobat*, *Trypanosomasp*, *Lewisi*, *Janseni*, *Leishmania_ infantum*, *Crithidia*, *Leishmania*, *Minasense*, *Dionisii* e *Bodonidae*. As distâncias dentro desse cluster variaram de 0.5 a 1.0, sugerindo que algumas dessas espécies podem estar mais próximas de certos tripanosomas do que de outras dentro do próprio grupo. Isso pode indicar a necessidade de uma análise mais detalhada para melhor compreensão das relações evolutivas dessas espécies.

DOI: 10.5540/03.2026.012.01.0317

Tabela 1: Espécies por clusters.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cascavelli	*Marinkellei*	*Rangeli*_A	*Neobodo_Designis*	*Neobat*
	*T. cruzi*TcBat	*Rangeli*_E	*Parabodo_Caudatus*	*Trypanosomasp*
	*T. cruzi*_IV	*Rangeli*_B	*Parabodo*	*Lewisi*
	*T. cruzi*_I	*Rangeli*_D		*Janseni*
	*T. cruzi*_II			*Leishmania_infantum*
	*T. cruzi*TcI			*Crithidia*
	*T. cruzi*TcIa			*Leishmania*
				Minasense
				Bodonidae
				Dionisii

Os graus de pertinência (Figura 1) reforçam a estrutura dos clusters formados. Espécies dentro de grupos bem definidos, como *Cascavelli* (Cluster 1), *T. cruzi* (Cluster 2) e *Rangeli* (Cluster 3), apresentam altos valores de pertinência, indicando uma classificação consistente. No Cluster 4, composto por espécies de vida livre, a pertinência também é elevada, refletindo sua distinção genética em relação aos parasitas. Já no Cluster 5, algumas espécies exibem pertinências intermediárias, sugerindo que podem compartilhar características genéticas com mais de um grupo. Isso pode indicar transição evolutiva, variação genética interna ou a necessidade de refinamento na definição desse cluster.

4 Conclusão

A abordagem híbrida (*fuzzy c-means* + clusterização hierárquica) mostrou-se promissora para clusterização envolvendo espécies Kinetoplastea com sequências não alinhadas. A estratégia baseada em *k-mers* evitou alinhamentos complexos, identificando cinco clusters geneticamente distintos, validados por alta correlação cophenética (0,96), FPC 0.6787 e especialistas. A análise da métrica de inercia agregadas a clusterização, também contribuiu para a melhor escolha de distribuição das espécies dentro dos clusters.

Sugere-se aplicar o método a outros gêneros, explorando diferentes técnicas para a construção da matriz de distâncias, como cadeias de Markov e integral fuzzy. Além disso, recomenda-se incorporar métricas adicionais para a validação dos clusters e aprimorar a visualização da distribuição das espécies nos grupos.

Referências

[1] L. E. Aik, T. W. Choon e M. S. Abu. "K-means algorithm based on flower pollination algorithm and calinski-harabasz index". Em: **Journal of Physics: Conference Series**. Vol. 2643. 1. IOP Publishing. 2023, pp. 012019–012026.

[2] J. Azé, C. Sola, J. Zhang, F. L. Marin, M. Yasmin, R. Siddiqui, K. Kremer, D. Van Soolingen e G. Refrégier. "Genomics and machine learning for taxonomy consensus: the Mycobacterium tuberculosis complex paradigm". Em: **PloS one** 10.7 (2015), e0130912.

[3] J. C. Bezdek. **Pattern recognition with fuzzy objective function algorithms**. Springer Science & Business Media, 2013.

[4] V. H. Borba, C. Martin, J. R. M. Silva, Sa. C. C. Xavier, F. L. de Mello e A. M. Iñiguez. "Machine learning approach to support taxonomic species discrimination based on helminth collections data". Em: **Parasites and Vectors** 14 (2021). ISSN: 1756-3305. DOI: 10.1186/s13071-021-04721-6.

[5] S. Derkarabetian, S. Castillo, P. K Koo, S. Ovchinnikov e M. Hedin. "A demonstration of unsupervised machine learning in species delimitation". Em: **Molecular phylogenetics and evolution** 139 (2019), p. 106562.

[6] G. A. Garcia, A. R. Lord, L. M. B. Santos, T. N. Kariyawasam, M. R. David, D. C. Lima, A. T. Ferreira, M. G. Pavan, M. T. Sikulu-Lord e R. M. Freitas. "Rapid and Non-Invasive Detection of Aedes aegypti Co-Infected with Zika and Dengue Viruses Using Near Infrared Spectroscopy". Em: **Viruses** 15.1 (2022), pp. 11–24.

[7] D.H. Le. "Machine learning-based approaches for disease gene prediction". Em: **Briefings in Functional Genomics** 19.5-6 (2020), pp. 350–363. ISSN: 2041-2657. DOI: 10.1093/bfgp/elaa013.

[8] D. M. Lopez, F. L. de Mello, C. M. Giordano Dias, P. Almeida, M. Araújo, M. A. Magalhães, G. S. Gazeta e R. P. Brasil. "Evaluating the Surveillance System for Spotted Fever in Brazil Using Machine-Learning Techniques". Em: **Frontiers in Public Health** 5 (2017), pp. 323–332. ISSN: 2296-2565. DOI: 10.3389/fpubh.2017.00323.

[9] C. Martin, N. N. Diaz, J. Ontrup e T. W. Nattkemper. "Hyperbolic SOM-based clustering of DNA fragment features for taxonomic visualization and classification". Em: **Bioinformatics** 24.14 (2008), pp. 1568–1574.

[10] M. C. Monard e J. A. Baranauskas. "Conceitos sobre aprendizado de máquina". Em: **Sistemas inteligentes-Fundamentos e aplicações** 1.1 (2003), pp. 89–114.

[11] F. Murtagh e P. Contreras. "Algorithms for hierarchical clustering: an overview". Em: **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery** 2.1 (2012), pp. 86–97.

[12] L. S. Ochi, C. R. Dias e S. S. F. Soares. "Clusterização em mineração de dados". Em: **Instituto de Computação-Universidade Federal Fluminense-Niterói** 1 (2004), p. 46.

[13] R. Resmini, A. Conci, T. B. Borchartt, R. C. F. de Lima, A. A. Montenegro e C. A. Pantaleão. "Diagnóstico precoce de doenças mamárias usando imagens térmicas e aprendizado de máquina". Em: **Revista brasileira de Contabilidade e Gestão** 1.1 (2012), pp. 55–67.

[14] R. R. Sokal e F. J. Rohlf. "The comparison of dendrograms by objective methods". Em: **Taxon** (1962), pp. 33–40.

[15] A. Zielezinski, S. Vinga, J. Almeida e W. M. Karlowski. "Alignment-free sequence comparison: benefits, applications, and tools". Em: **Genome biology** 18 (2017), pp. 1–17.

