Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Aplicação de Lógica *Fuzzy* na Taxonomia e Filogenia de *Leishmania spp.*

Antonio R. A. Pereira[1], Marcello G. Teixeira[2], Flávio L. de Mello[3]
UFRJ, Rio de Janeiro, RJ

Resumo. A lógica *fuzzy*, que lida com incerteza e imprecisão, é aplicada neste estudo para analisar a filogenia e a taxonomia do gênero *Leishmania*, buscando uma modelagem mais precisa e flexível da diversidade biológica. Foram coletadas e preparadas sequências de DNA das espécies, seguidas da aplicação do algoritmo *fuzzy c-means* para uma análise detalhada. Os resultados revelaram agrupamentos com coeficientes de partição *fuzzy* variados, sendo o de sete *clusters* o mais próximo da representação taxonômica tradicional, desafiando expectativas geográficas e taxonômicas. A integração da lógica *fuzzy* demonstrou ser promissora, oferecendo uma abordagem flexível e valiosa para entender as relações evolutivas e a organização taxonômica das leishmanias.

Palavras-chave. Aprendizado de Máquina, Aprendizado não-supervisionado, Clusterização, Filogenia, Taxonomia, Lógica *Fuzzy*

1 Introdução

A lógica *fuzzy* é um conceito que busca modelar de forma mais precisa, mesmo lidando com a imprecisão inerente, a maneira como respondemos a questões do cotidiano que não se encaixam facilmente em categorias binárias como "sim"ou "não". Ela utiliza palavras em vez de números para expressar os valores verdadeiros, tornando-se especialmente útil em situações onde a precisão não é claramente definida.

Para ilustrar, considere questões como determinar se uma pessoa com 1,68 metros é alta ou baixa, se alguém com 20 anos é jovem ou adulto, ou avaliar se um vazamento de petróleo em uma praia é pequeno, médio ou grande. Nessas situações, a lógica *fuzzy* permite expressar as respostas de forma mais refinada, usando termos linguísticos em vez de valores numéricos.

O conceito de lógica *fuzzy* foi estabelecido pelo Dr. Lotfi Zadeh em seu trabalho de 1964 [21], estabelecendo as bases para um campo de estudo que encontrou aplicações em diversos domínios do conhecimento. Em contraste com a lógica booleana, que lida com valores binários (verdadeiro ou falso), a lógica *fuzzy* permite a utilização de valores lógicos que variam em um intervalo contínuo de $[0, 1]$.

Portanto, a lógica *fuzzy* oferece uma abordagem mais flexível e adaptável para lidar com a incerteza e a imprecisão presentes em muitos problemas do mundo real. Ela se destaca especialmente em contextos onde as respostas não podem ser facilmente reduzidas a simples afirmativas binárias.

A grande motivação para o desenvolvimento deste trabalho remonta a 1758, quando o cientista sueco Carl Nilsson Linnaeus (Lineu) propôs um critério de classificação de seres vivos, considerando características estruturais e anatômicas. Porém, Lineu era criacionista e acreditava que o número de espécies era fixo e imutável, assim como foram criados pela entidade divina [9]. O sistema

[1] antoniorevail@ppgi.ufrj.br
[2] marcellogt@dcc.ufrj.br
[3] fmello@poli.ufrj.br

criado por Lineu é rígido e não considera que os seres vivos estejam em constante evolução. A observação do mundo natural precisa admitir incerteza, instabilidade e mutabilidade porque os fenômenos naturais dificilmente apresentam contornos bem definidos, uma vez que se encontram em constantes mudanças. Dessa forma, ao integrar a lógica *fuzzy* na filogenia e taxonomia, conseguimos uma modelagem mais fiel à complexidade da vida, capturando a diversidade biológica de maneira mais precisa e flexível. Isso contribui para uma compreensão mais robusta das relações evolutivas e organização taxonômica dos seres vivos.

Para validar nossa hipótese, optamos por selecionar como nosso modelo inicial um táxon parasita de grande importância médica, o gênero *Leishmania*. Os morfotipos foram descritos em 1903 pelo médico britânico Ronald Ross, sendo o nome uma homenagem ao patologista escocês William Boog Leishman [14], que, em maio do mesmo ano, descreveu pequenos corpos ovais encontrados no baço de um soldado morto após apresentar sintomas como febre alta, disenteria crônica e caquexia [15]. Para uma compreensão abrangente do gênero, várias tentativas de taxonomia foram realizadas, considerando diversas características das Leishmânias, dentre elas destacam-se [7, 13, 16, 18, 19]. Mais recentemente, [4] dividiram o gênero em duas grandes linhagens filogenéticas, *Euleishmania* e *Paraleishmania*, sendo a primeira de maior importância médica e composta por quatro subgêneros: *Leishmania*, *Viannia*, *Sauroleishmania* e o complexo *L. enriettii*, posteriormente identificado como *Mundinia*. Para este trabalho, consideramos somente espécies dos subgêneros *Leishmania*, *Viannia* e *Mundinia*.

Com o desenvolvimento das tecnologias de sequenciamento genético, aplicadas nos estudos em filogenia molecular que analisam diferenças moleculares, principalmente nas sequências de DNA, permite obter informações sobre as relações evolutivas entre organismos. Como consequência, revisões em taxonomias têm sido cada vez mais presentes na literatura. O que antes era feito observando aspectos morfológicos, atualmente consideram-se aspectos moleculares.

À medida que a quantidade de dados moleculares aumenta, enfrentamos desafios em sua manipulação devido ao volume considerável. Para lidar com essa complexidade, os algoritmos de aprendizado de máquina têm sido amplamente adotados em estudos moleculares, oferecendo não apenas uma abordagem mais rápida para processamento, mas também resultados de alta qualidade. Neste estudo, foi utilizado um método de aprendizado não supervisionado baseado na lógica *fuzzy*, conhecido como *fuzzy c-means*, para explorar e analisar os dados moleculares de forma mais eficaz. Esse método não apenas realiza a segmentação dos dados em *clusters*, mas também fornece o grau de pertinência de cada espécie em relação ao *cluster* correspondente. Essa abordagem permite uma análise mais refinada da distribuição das espécies nos agrupamentos identificados, fornecendo *insights* valiosos sobre a relação de cada espécie com seu respectivo *cluster*.

2 Pré-processamento

O conjunto de dados foi coletado no *website* [10] do Instituto Nacional de Saúde americano. Sequências de DNA dos cromossomos 10, 28 e 31 das leishmanias (*L. aethiopica, L. arabica, L. braziliensis, L. donovani, L. enriettii, L. infantum, L. major, L. mexicana, L. panamensis*) e *L. peruviana* foram adquiridas em formato $.fasta$. Os cromossomos 28 e 31 foram escolhidos por conterem os genes $hsp70$ e $6pg$, característicos da espécie, e o cromossomo 10 foi escolhido aleatoriamente.

A etapa subsequente envolveu o alinhamento das sequências, visando uniformizar a quantidade de bases em todas elas. Dadas as características dos dados, optou-se por utilizar um programa capaz de realizar alinhamentos múltiplos em sequências extensas. O Mauve é um sistema que não apenas executa alinhamentos tradicionais de múltiplas sequências, mas também integra a análise de eventos evolutivos em larga escala [5]. O Mauve foi a escolha ideal, atendendo às especificidades do conjunto de dados.

DOI: 10.5540/03.2025.011.01.0350

Devido à limitação de apenas 10 sequências, uma para cada espécie, optou-se por gerar artificialmente um conjunto expandido de sequências. Esse processo envolveu a introdução de mutações nas sequências base. As mutações genéticas podem ser classificadas em três tipos, sendo a substituição uma delas, caracterizada pela troca de um ou mais pares de bases na sequência de nucleotídeos [12].

O modelo de substituição Jukes-Cantor, adotado neste estudo, pressupõe que as frequências de equilíbrio de todas as bases são iguais, e as mudanças nucleotídicas ocorrem na mesma taxa em todos os sítios [17].

Devido aos alinhamentos, observou-se a presença de *gaps* $(-)$ ou a letra N em alguns sítios. Considerando que esses elementos representam falta de informação, optou-se por substituir todos os *gaps* por N.

Na construção do algoritmo de substituição, considerou-se a ocorrência de mutações, na natureza, em até 1% do genoma. Sorteou-se um valor de porcentagem entre $0,1$ e 1, que representa quantos sítios da sequência de entrada sofreriam substituições. Estabelecidos os sítios que sofreriam as mutações, o próximo passo foi, reconhecendo a base que está no sítio, substituir de maneira aleatória por outra base. Foram produzidas um total de quarenta sequências baseadas neste modelo.

3 Vetorização

A vetorização desempenha um papel crucial neste trabalho, pois é por meio dela que as sequências de bases nitrogenadas, Adenina (A), Citosina (C), Guanina (G) e Timina (T), são convertidas em vetores numéricos.

Foram realizadas algumas outras vetorizações [11] para melhor representar o conjunto de dados. A abordagem que mais se adequou às condições deste trabalho é conhecida como vetorização por número atômico [8]. Nesta técnica, as bases A, C, G, T e N foram convertidas em sequências numéricas, correspondendo ao número total de prótons em cada nucleotídeo: $A = 70$, $T = 66$, $C = 58$, $G = 78$, $N = 0$. De forma que uma sequência formada pelos nucleotídeos $[AGCCTTGN]$ pode ser representada pelo vetor [70, 78, 58, 58, 66, 66, 78, 0]. Essa vetorização foi empregada na obtenção dos resultados apresentados neste artigo. A principal vantagem desta abordagem em relação às outras é o tamanho do arquivo resultante (1,1 GB), contendo os vetores que representam as sequências após a vetorização.

4 Algoritmo de Agrupamento Baseado em Lógica *Fuzzy*

Clusterização é um processo de agrupamento de dados não supervisionado em que a existência de padrões na base de dados é percebida pelo próprio algoritmo, que verifica a similaridade entre esses dados e os agrupa, para obter alta homogeneidade dentro dos grupos e alta heterogeneidade entre os grupos.

O algoritmo *fuzzy c-means* (FCM) é um algoritmo baseado em lógica *fuzzy* que considera o grau de pertinência de cada ponto do conjunto de dados em relação aos diversos agrupamentos. Foi proposto primeiramente por [6] e posteriormente generalizado por [2], que define este algoritmo como um programa que gera partições difusas e protótipos para qualquer conjunto de dados numéricos. De forma geral, o FCM tenta dividir os dados em conjuntos minimizando a função objetivo (1).

$$J = \sum_{i=1}^{n} \sum_{j=1}^{p} \mu_{ij}^{m} d(x_i : c_j)^2 \tag{1}$$

DOI: 10.5540/03.2025.011.01.0350

onde μ_{ij} é o grau de pertinência da amostra x_i ao j-ésimo *cluster*, n é o número de instâncias, p é o número de *clusters* considerados no FCM (que deve ser decidido antes da execução), $m > 1$ é o parâmetro fuzzyficador (usualmente a escolha é por um número irracional no intervalo [1,25;2] [3]), x_i é um vetor de dados de treinamento onde $i = 1, 2, ..., n$ e representa um atributo do dado, c_j, onde $j = 1, 2, ..., p$, é o centro de um agrupamento *fuzzy* e $d(x_i : c_j)$ é a distância euclidiana entre x_i e c_j.

A distinção entre algoritmos de clusterização, como o *k-means*, OPTICS, e os algoritmos de clusterização hierárquica, em comparação com o *fuzzy c-means*, reside no fato de que este último permite que os pontos de dados pertençam a vários *clusters* simultaneamente. No entanto, uma desvantagem associada a essa abordagem é a necessidade de especificar previamente o número de *clusters* desejados. Essa questão foi, em parte, superada pela execução repetida do algoritmo, observando a qualidade dos *clusters* através do coeficiente de partição *fuzzy*.

Foi empregada uma implementação em *Python*, utilizando o pacote *skfuzzy*, disponível no *website* [20].

5 Resultados e Discussões

A taxonomia discutida no trabalho de [1] aborda os subgêneros atuais de *Leishmania*, *Viannia* e *Mundinia*. Optou-se por focar exclusivamente nas espécies que parasitam mamíferos, excluindo do escopo deste trabalho aquelas que infectam répteis, pertencentes ao subgênero *Sauroleishmania*

No algoritmo *fuzzy c-means*, foram empregadas cinquenta sequências com a vetorização por número atômico. O FCM tem como base a associação de um objeto a todos os *clusters*, usando a função de pertinência. A validação dos *clusters* é dada pelo coeficiente de partição *fuzzy*, assumindo valores no intervalo $[0, 1]$, sendo 1 o melhor. O programa foi aplicado variando entre 2 e 10 *clusters*.

A Tabela 1 mostra que os valores de coeficiente de partição *fuzzy* variaram entre 0,40785 e 0,87481. Assim, no agrupamento com 10 *clusters*, foi observado o valor mais alto do coeficiente de partição *fuzzy*, atingindo a separação máxima. Porém, agrupamentos com dez *clusters* não são ideais porque significaria, por exemplo, colocarmos as *L. donovani* e *L. infantum* em *clusters* diferentes quando a literatura as coloca juntas.

Tabela 1: Quantidade de clusters e coeficiente de partição

Quantidade de clusters	**Coeficiente de Partição**
2	0,52483
3	0,49749
4	0,43940
5	0,56039
6	0,61419
7	0,68551
8	0,44459
9	0,40785
10	0,87481

Na primeira tabela da Tabela 2, temos as espécies distribuídas nos sete *clusters*. As espécies dos *clusters* 1 e 2 aparecem agrupadas com espécies diferentes do que ocorre na taxonomia convencional, mantendo, porém, a coerência em relação aos subgêneros.

A segunda tabela da Tabela 2 apresenta o *cluster* onde as espécies de *L. aethiopica* estão agrupadas, isto é verificado pelo maior valor das linhas. Os valores menores são os de pertinência da espécie em relação aos outros *clusters*.

Tabela 2: Agrupamento com 7 clusters

Agrupamento com 7 clusters							
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**
	L. braziliensis *L. panamensis*	*L. major*	*L. peruviana*	*L. donovani* *L. infantum*	*L. enriettii*	*L. aethiopica*	*L. arabica*
Espécies	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**
L. aethiopica base	0,00283538	0,00189467	0,02010416	0,00354264	0,00319455	0,98293219	0,003506206
L. aethiopica 1	0,015925914	0,010782328	0,011964421	0,019987718	0,017907856	0,008354268	0,015966384
L. aethiopica 2	0,003665983	0,002461349	0,002733374	0,00459825	0,006147011	0,977843394	0,00455464
L. aethiopica 3	0,008519494	0,005738579	0,006370907	0,010668743	0,009644417	0,94835454	0,010551261
L. aethiopica 4	0,010558704	0,007121699	0,007904724	0,01321117	0,011903164	0,936235029	0,01306509

No gráfico da Figura 1, cada barra representa uma espécie e os *clusters* são representados pelas cores. A distribuição das cores nas barras representa a pertinência de cada espécie nos *clusters*. O eixo vertical varia no intervalo [0, 1] e representa o valor das pertinências para cada *cluster*.

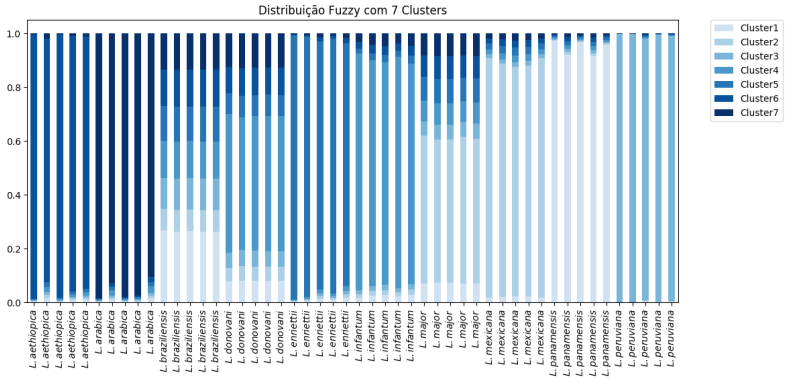

Figura 1: Gráfico com 7 *clusters*. Fonte: Elaborada pelo autor.

O agrupamento com sete *clusters*, além de apresentar o segundo maior valor de coeficiente de partição *fuzzy*, é a clusterização mais próxima da representação taxonômica clássica descrita para os complexos do gênero. O resultado com sete *clusters fuzzy* também manteve a divisão em sete complexos, porém com modificações no nível de espécies que formam os complexos. Por exemplo, no *cluster* 2 da Tabela 1, as leishmanias *L. major* e *L. mexicana* aparecem juntas no mesmo *cluster*, diferente do que mostra a taxonomia de [1]. A *L. major* é uma espécie do Novo Mundo, enquanto a espécie *L. mexicana* é do Velho Mundo. A hipótese é que elas tenham um ancestral comum e habitavam a mesma área, sofrendo isolamento geográfico e se separando. O isolamento geográfico também pode explicar o fato de *L. braziliensis* agrupar com *L. peruviana* e se separar de *L. panamensis*. Neste caso, *L. braziliensis* e *L. peruviana* podem ter compartilhado o mesmo ancestral, que se dividiu em populações separadas por barreiras geográficas. Desse ponto em diante, as populações podem ter evoluído independentemente em seus respectivos habitats, acumulando diferenças genéticas ao longo do tempo. Por sua vez, *L. panamensis* pode ter permanecido isolada em uma área geográfica diferente, resultando em uma divergência genética mais significativa.

6 Considerações Finais

A escolha pelo algoritmo FCM, que se baseia em conjuntos *fuzzy*, deve-se à vantagem de expressar o tipo de situação em que um objeto compartilha similaridade com vários grupos. Sendo assim, o FCM associa cada indivíduo parcialmente a todos os grupos.

Os graus de pertinência indicam que as espécies apresentam algum grau de pertencimento aos demais complexos. Se, por um lado, isso representa uma origem comum entre todas as espécies, o que é esperado, por outro, demonstra que a definição de uma espécie não pode obedecer a critérios tão rígidos, mas sim considerar graus de identidade entre diferentes espécies, uma representação mais fiel do contraste e do ininterrupto fenômeno de evolução das espécies. Com esses valores, podemos ainda quantificar a proximidade ou o afastamento dessas espécies entre si e em relação a cada complexo. A aplicação de um algoritmo baseado em lógica *fuzzy* revelou padrões intrigantes que desafiam as expectativas baseadas na distribuição geográfica e na taxonomia tradicional.

Sendo assim, sugere-se o aprofundamento dos estudos sobre a utilização do FCM para a classificação de espécies.

Referências

[1] M. Akhoundi, T. Downing, J. Votỳpka, K. Kuhls, J. Lukeš, A. Cannet, C. Ravel, P. Marty, P. Delaunay, M. Kasbari et al. "Leishmania infections: Molecular targets and diagnosis". Em: **Molecular aspects of medicine** 57 (2017), pp. 1–29. DOI: 10.1016/j.mam.2016.11.012.

[2] J. C. Bezdek, R. Ehrlich e W. Full. "FCM: The fuzzy c-means clustering algorithm". Em: **Computers & geosciences** 10.2-3 (1984), pp. 191–203. DOI: 10.1016/0098-3004(84)90020-7.

[3] E. Cox. **Fuzzy modeling and genetic algorithms for data mining and exploration**. Elsevier, 2005. DOI: 10.1016/B978-0-12-194275-5.X5000-2.

[4] E. Cupolillo, E. Medina-Acosta, H. Noyes, H. Momen e G. Grimaldi. "A revised classification for Leishmania and Endotrypanum". Em: **Parasitology today** 16.4 (2000), pp. 142–144.

[5] A. C. E. Darling, B. Mau, F. R. Blattner e N. T. Perna. "Mauve: multiple alignment of conserved genomic sequence with rearrangements". Em: **Genome research** 14.7 (2004), pp. 1394–1403.

[6] J. C. Dunn. "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". Em: **Journal of Cybernetics** 3.3 (1973), pp. 32–57. DOI: 10.1080/01969727308546046.

[7] S. M. Gossage, M. E. Rogers e P. A. Bates. "Two separate growth phases during the development of Leishmania in sand flies: implications for understanding the life cycle". Em: **International journal for parasitology** 33.10 (2003), pp. 1027–1034.

[8] T. Holden, R. Subramaniam, R. Sullivan, E. Cheung, C. Schneider, G. Tremberger Jr., A. Flamholz, D.H. Lieberman e T.D. Cheung. "ATCG nucleotide fluctuation of Deinococcus radiodurans radiation genes". Em: 6694 (2007), p. 669417.

[9] V. Klepka e M.J. Corazza. "O essencialismo na classificação de Lineu e a repercussão dessa controvérsia na Biologia". Em: **História da Ciência e Ensino: construindo interfaces** 18 (2018), pp. 73–110.

[10] NCBI. **National Center for Biotechnology Information**. Online. https:www.ncbi.nlm.nih.gov.

[11] A. R. A. Pereira. "Classificação de complexos de Leishmania usando algoritmos de agrupamentos". Dissertação de mestrado. PPGI/UFRJ, 2022.

[12] B. A. Pierce. **Genética: um enfoque conceitual. 5ª edição**. Rio de Janeiro: Guanabara Koogan, 2016.

[13] J.A. Rioux, G. Lanotte, E. Serres, F. Pratlong, P. Bastien e J. Perieres. "Taxonomy of Leishmania. Use of isoenzymes. Suggestions for a new classification". Em: **Annales de parasitologie humaine et comparee** 65.3 (1990), pp. 111–125.

[14] R. Ross. "Further notes on Leishman's bodies". Em: **British medical journal** 2.2239 (1903), p. 1401.

[15] R. Ross. "Note on the bodies recently described by Leishman and Donovan". Em: **British medical journal** 2.2237 (1903), p. 1261.

[16] V.M. Safj'anova. "The problem of taxonomy with Leishmania". Em: **Ser Protozool Sov Acad Sci Leningr** 7 (1982), pp. 5–109.

[17] H. Schneider. **Métodos de análise filogenética: um guia prático**. Holos, 2003.

[18] J.R. Stevens, H.A. Noyes, C.J. Schofield e W. Gibson. "The molecular evolution of Trypanosomatidae". Em: **Parasitology today** 48 (2001), pp. 1–53.

[19] V. Thomaz-Soccol, G. Lanotte, J.A. Rioux, F. Pratlong, A. Martini-Dumas e E. Serres. "Monophyletic origin of the genus Leishmania Ross, 1903". Em: **Annales de parasitologie humaine et comparée** 68 (1993), pp. 107–107.

[20] J. Warner. **Scikit-fuzzy**. Online. https:pythonhosted.org/scikit-fuzzy.

[21] L.A. Zadeh. "Fuzzy sets". Em: **Information and control** 8.3 (1965), pp. 338–353.

