

# Aplicação de métodos de aprendizado de máquina não-supervisionado em dados de metilação de DNA de pacientes de COVID-19

Orientadora: Profa. Samara Kiihl  
Aluno: Guilherme Pereira de Freitas

28 de julho de 2022

## Resumo

Estudos recentes vêm demonstrando a importância da metilação de DNA, um marcador epigenético importante, na regulação da expressão gênica. Métodos de agrupamento ou de aprendizado de máquina não-supervisionado são utilizados para extrair informações para diagnóstico precoce e tratamentos a partir de dados de alta dimensão dos estudos epigenômicos. Neste projeto, iremos aplicar métodos de agrupamento em dados de metilação de DNA de 407 pacientes de COVID-19, onde 194 (47.7%) foram diagnosticados com sintomas leves e 213 (52.3%) com sintomas severos, com o objetivo de encontrar grupos que podem ter sido formados por influência dessa doença.

**Palavras-chave:** bioinformática, metilação de DNA, aprendizado de máquina não supervisionado, métodos de agrupamento, COVID-19, EWAS

## 1 Introdução

Embora todas as células de um organismo apresentem, essencialmente, o mesmo conteúdo genético, suas funções e particularidades se dão por meio do regulamento da expressão gênica. Tal regulamento ocorre por meio de mecanismos epigenéticos, como a metilação do DNA, modificação de histonas e outros processos mediados por RNA, que influenciam principalmente a expressão gênica a nível de transcrição [1].

O mecanismo estudado foi a metilação da citosina (5mC), que acontece em áreas específicas de regulação, como regiões promotoras ou de heterocromatina. Esse fenômeno pode modificar, significativamente, a expressão temporal e espacial dos genes e a remodelação da cromatina [2]. Em mamíferos, as citosinas metiladas estão restritas às CpGs (citosina-fosfato-guanina), onde elas antecedem uma guanina (G) na direção de 5'. Vale lembrar que o DNA é formado por 4 nucleotídeos e portanto existem 16 possibilidades para se formar um par em sequência, o que ajuda a identificar as ilhas CpGs, pois estas apresentam uma frequência maior desse par (CG) quando comparadas com outras regiões arbitrárias [1].

## 2 Banco de dados

O banco de dados utilizado foi retirado do estudo de associação epigenômica ampla (EWS - *epigenome-wide association study*) de COVID-19 realizado por [3]. Os dados estão disponíveis no repositório público *The Gene Expression Omnibus* (GEO), sob código de acesso GSE168739. Trata-se de 407 pacientes de COVID-19 sem comorbidades e com idade máxima de 61 anos, onde 194 (47.7%) estavam com sintomas leves e 213 (52.3%) estavam com sintomas graves. Os dados foram coletados através da plataforma *Infinium MethylationEPIC BeadChip*, totalizando em aproximadamente 850 mil ilhas CpGs para cada indivíduo.

### 3 Pré-processamento

O fluxo de pré-processamento foi feito seguindo o passo a passo descrito no artigo "A cross-package Biodonductor workflow for analysing methylation array data" [4], por meio das ferramentas dispostas no pacote Bioconductor [5], disponíveis para a linguagem R [6]. Os algoritmos são aplicados na matriz de p-valores.

A matriz de p-valores é obtida comparando-se a distribuição das intensidades, para cada par de indivíduos e ilhas, com a distribuição do ruído de fundo (que por sua vez, foi calculado a partir das sondas de controle). Cada um dos ensaios (tipo I e tipo II) apresenta sua distribuição própria do ruído de fundo, bem como a intensidade de metilação dos indivíduos. Como exemplo, tomemos um indivíduo qualquer presente no banco. O primeiro passo é filtrar as sondas de controle, em cada um dos tipos de ensaios, e em seguida obter os parâmetros de três distribuições normais  $\mathcal{N}(2\mu, 2\sigma^2)$  (Red, Green, Green+Red), onde  $\mu$  é a mediana e  $\sigma^2$  é o desvio absoluto mediano das intensidades para essas sondas. Após isso, devemos obter a intensidade de metilação total do indivíduo em cada ilha e calcular a probabilidade de cada uma dessas intensidades ser uma amostra da distribuição normal obtida no início [7].

Sem entrar em muitos detalhes, as etapas de pré-processamento consistiram em remover amostras de baixa qualidade, aplicar a normalização quantílica [8] e remover sondas de CpGs referentes ao sexo do paciente e/ou com reatividade cruzada [9].

Devido a limitações computacionais para realizar as etapas de pré-processamento, esse trabalho usou recursos do Centro Nacional de Processamento de Alto Desempenho em São Paulo (CENAPAD-SP). Todos os códigos do projeto podem ser acessados no repositório dessa pesquisa presente no Github [10].

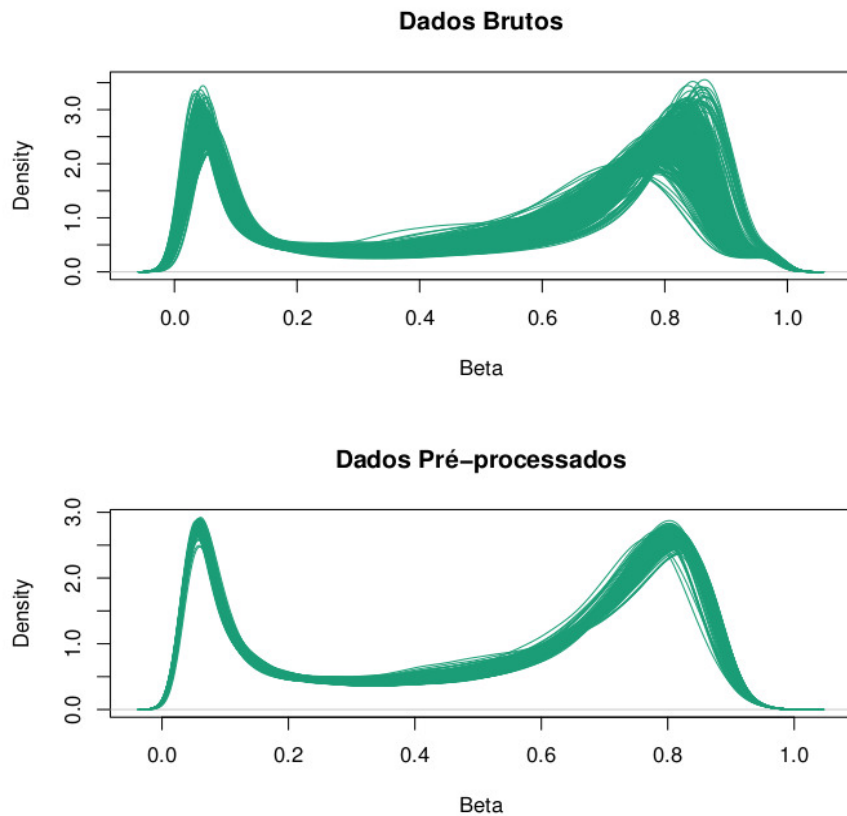


Figura 1: Densidade das taxas de metilação antes e depois do pré-processamento

## 4 Metodologia

O aprendizado de máquina não-supervisionado também é conhecido como análise de cluster, ou análise de agrupamento. Uma das maiores diferenças entre aprendizado de máquina supervisionado e não-supervisionado está na falta de dados de treinamento para a última, bem como a falta de um target para tal. Os pré-requisitos para aplicar as técnicas de agrupamento se dão na escolha das variáveis, hiperparâmetros e tipo de distância adotada [11].

Com a matriz final dos betas pré-processados, podemos calcular a matriz de dissimilaridade entre os indivíduos por meio da distância euclidiana, dada pela fórmula  $D(X, Y) = \sqrt{\sum_i^n (x_i - y_i)^2}$ , onde  $n$  é o número total de CpGs e  $X$  e  $Y$  são os vetores de betas de dois indivíduos. Cada um dos métodos foi aplicado na matriz de distâncias completa e em duas componentes obtidas através do método Uniform Manifold Approximation and Projection (UMAP) [12].

Para avaliar a formação de clusteres, em cada um dos métodos, utilizou-se as técnicas de Silhueta [13] e Gráfico de Elbow [14].

- Silhueta: Dado um conjunto de clusteres  $\Lambda$ , temos que a silhueta da observação  $i$  presente no cluster  $\lambda_k$  é dada por  $s_{i\lambda_k} = \frac{b_i - a_i}{\max(b_i, a_i)}$ , onde  $a_i$  é a dissimilaridade de  $i$  com relação aos elementos do cluster  $\lambda_k$  (que o contém) e  $b_i$  é a menor dissimilaridade de  $i$  com relação aos elementos de outro cluster  $\lambda$ , ou seja,  $b_i = \min_{\lambda \neq \lambda_k} d(i, \lambda)$ .

- Método de Elbow: O Método de Elbow é uma curva construída a partir da Soma de Quadrados Intra-cluster, ou inércia, cuja fórmula é dada por  $WSS = \sum_{k=1}^K \sum_{i \in S_k} \sum_{j=1}^p (x_{ij} - \bar{x}_{kj})^2$ . O melhor número de clusteres é obtido no ponto de maior inclinação da curva.

Os métodos de agrupamento utilizados foram K-Médias [15], PAM - Partition Around Medoids [16], Agrupamento Hierárquico [17] e Agrupamentos Baseados em modelos [18].

## 5 Resultados

Diferentes matrizes de dissimilaridade foram submetidas em cada um dos métodos comentados anteriormente, com o objetivo de verificar se existem dois grupos significativos possivelmente formados por efeito da Covid-19.

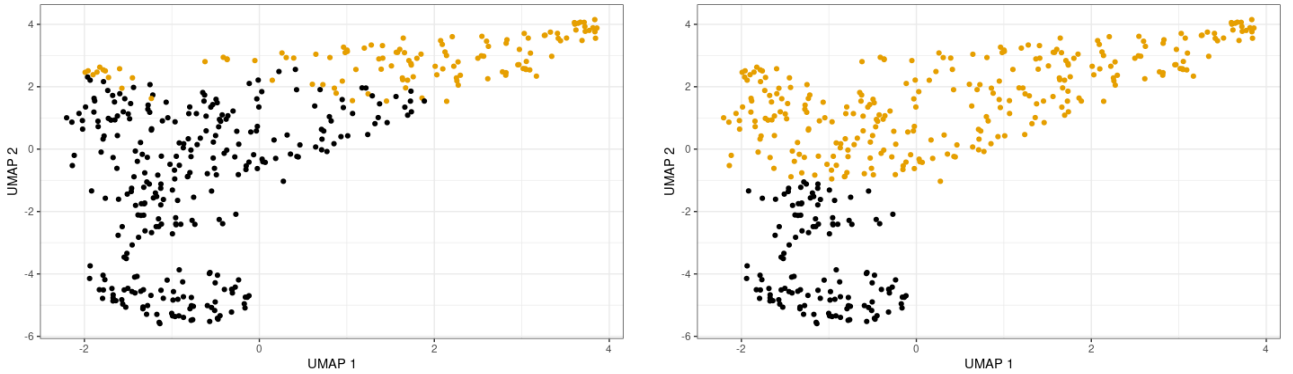


Figura 2: K-Médias: Agrupamentos encontrados através da matriz de distâncias completa (à esquerda) e através da matriz de distâncias das componentes UMAP (à direita).

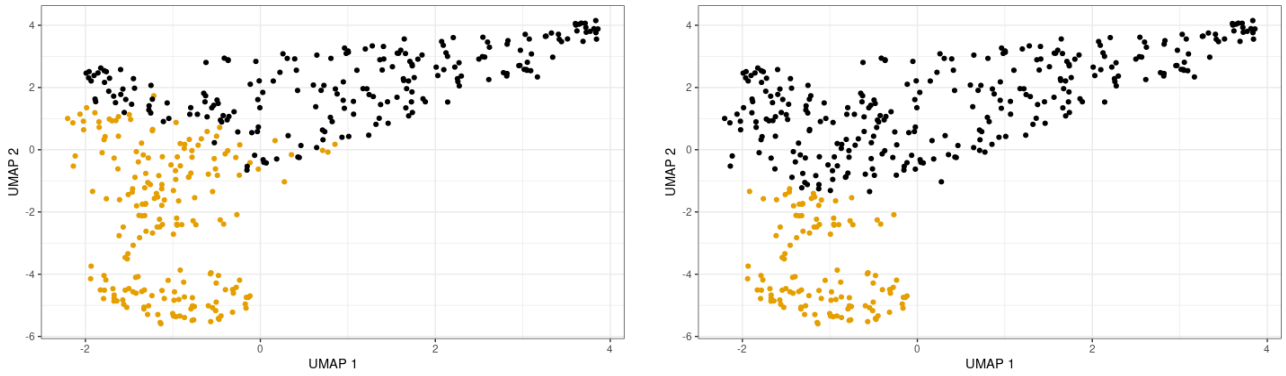


Figura 3: PAM: Agrupamentos encontrados através da matriz de distâncias completa (à esquerda) e através da matriz de distâncias das componentes UMAP (à direita).

De forma geral, os algoritmos PAM e K-Médias tiveram performances similares em termos de formação de grupos. Veja que os grupos formados lembram elipses, e por conta disso, pareceu interessante testar agrupamentos por Mistura de Modelos Gaussianos (GMM).

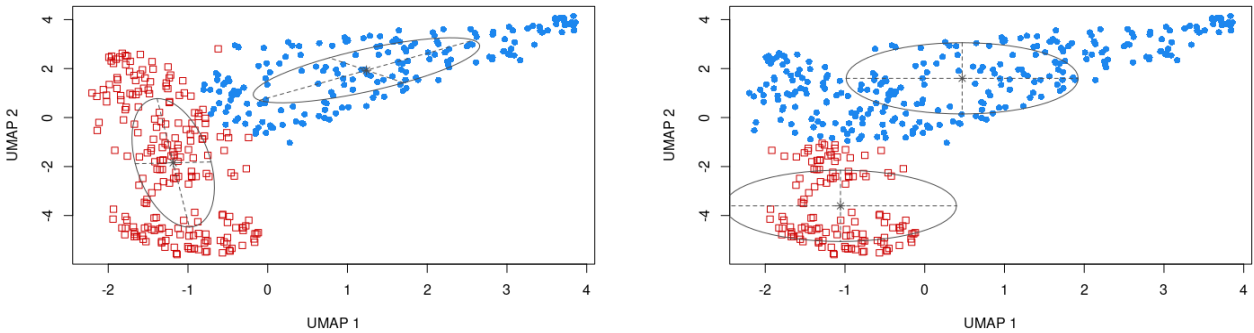


Figura 4: Mistura de Modelos Gaussianos: Agrupamentos encontrados através da matriz de distâncias das componentes UMAP. EVV (à esquerda) e EII (à direita).

## 6 Conclusão

Através dessa pesquisa, foi possível aprender diversos temas de biologia e bioinformática, bem como aprofundar em conceitos mais complexos de programação e métodos de aprendizado de máquina não-supervisionado. Foram exploradas diversas técnicas de agrupamentos para cada um dos métodos comentados e as conclusões finais ainda estão sendo escritas e sumarizadas.

De forma geral, podemos dizer que, utilizando de todas as CpGs que restaram após a etapa de pré-processamento, não foi possível detectar formação de grupos, pois a média da silhueta ficou muito pequena em todas as técnicas. No entanto, o clustering feito nas componentes do UMAP permitiu-nos encontrar 2 grupos significativos, cuja silhueta ficou próxima de 0,5. Devido à falta de acesso aos dados acerca da gravidade da COVID-19, não podemos checar se os grupos formados podem ter sido decorrentes do grau dessa doença.

## Referências

- [1] E R Gibney e C M Nolan. “Epigenetics and gene expression”. Em: *Heredity* 105.1 (mai. de 2010), pp. 4–13. DOI: 10.1038/hdy.2010.54. URL: <https://doi.org/10.1038/hdy.2010.54>.
- [2] Holger Heyn e Manel Esteller. “DNA methylation profiling in the clinic: applications and challenges”. Em: *Nature Reviews Genetics* 13.10 (set. de 2012), pp. 679–692. DOI: 10.1038/nrg3270. URL: <https://doi.org/10.1038/nrg3270>.
- [3] Manuel Castro de Moura et al. “Epigenome-wide association study of COVID-19 severity with respiratory failure”. Em: *EBioMedicine* 66 (abr. de 2021), p. 103339. DOI: 10.1016/j.ebiom.2021.103339. URL: <https://doi.org/10.1016/j.ebiom.2021.103339>.
- [4] Jovana Maksimovic, Belinda Phipson e Alicia Oshlack. “A cross-package Bioconductor workflow for analysing methylation array data”. Em: *F1000Research* 5 (jun. de 2016), p. 1281. DOI: 10.12688/f1000research.8839.1. URL: <https://doi.org/10.12688/f1000research.8839.1>.
- [5] W. Huber et al. “Orchestrating high-throughput genomic analysis with Bioconductor”. Em: *Nature Methods* 12.2 (2015), pp. 115–121. URL: <http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html>.
- [6] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2020. URL: <https://www.R-project.org/>.
- [7] Martin J. Aryee et al. “Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays”. Em: *Bioinformatics* 30.10 (jan. de 2014), pp. 1363–1369. DOI: 10.1093/bioinformatics/btu049. URL: <https://doi.org/10.1093/bioinformatics/btu049>.
- [8] Nizar Touleimat e Jörg Tost. “Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation”. Em: *Epigenomics* 4.3 (jun. de 2012), pp. 325–341. DOI: 10.2217/epi.12.21. URL: <https://doi.org/10.2217/epi.12.21>.
- [9] Yi-an Chen et al. “Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray”. Em: *Epigenetics* 8.2 (fev. de 2013), pp. 203–209. DOI: 10.4161/epi.23470. URL: <https://doi.org/10.4161/epi.23470>.
- [10] Guilherme Freitas. *Projeto de Iniciação Científica*. Fev. de 2022. URL: <https://github.com/GuilhermeFreitas09/IC>.
- [11] R. Gentleman e V. J. Carey. “Unsupervised Machine Learning”. Em: *Bioconductor Case Studies*. Springer New York, 2008, pp. 137–157. DOI: 10.1007/978-0-387-77240-0\_10. URL: [https://doi.org/10.1007/978-0-387-77240-0\\_10](https://doi.org/10.1007/978-0-387-77240-0_10).
- [12] Leland McInnes, John Healy e James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. arXiv: 1802.03426 [stat.ML].
- [13] Peter J. Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. Em: (1986).
- [14] Satoru Hayasaka. *How many clusters?* Fev. de 2022. URL: <https://towardsdatascience.com/how-many-clusters-6b3f220f0ef5>.
- [15] S. Lloyd. “Least squares quantization in PCM”. Em: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 129–137. DOI: 10.1109/TIT.1982.1056489.
- [16] Leonard Peter J - Kaufman Rousseeuw. *Finding Groups in Data*. John Wiley Sons, Inc., 2009.
- [17] K V Mardia, J M Bibby e J T Kent. *Multivariate analysis*. 1979, xv, 521 p. : ISBN: 0124712509 0124712525 0124712525. URL: <http://www.loc.gov/catdir/toc/els031/79040922.html>.
- [18] Charles Bouveyron et al. *Model-Based Clustering and Classification for Data Science*. Cambridge University Press, jun. de 2019. DOI: 10.1017/9781108644181. URL: <https://doi.org/10.1017/9781108644181>.