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

**Guilherme Pereira de Freitas (IMECC/UNICAMP)**, **Samara Flamini Kiihl (IMECC/UNICAMP)**

Financiamento: PIBIC/CNPq. Palavras-Chave: bioinformática, metilação de DNA, aprendizado de máquina não-supervisionado, métodos de agrupamento, COVID-19, EWAS.

## Introdução

Estudos recentes vêm demonstrando a importância da metilação de DNA, um marcador epigenético importante na regulação da expressão gênica [1]. Métodos de aprendizado de máquina não-supervisionado, como agrupamentos, são utilizados para extrair informações para diagnóstico precoce e tratamentos a partir de dados de alta dimensão dos estudos epigênicos. Neste projeto, aplicou-se métodos de agrupamento em dados de metilação de DNA de pacientes com COVID-19, com o objetivo de encontrar grupos que podem ter sido formados por influência dessa doença.

## Metodologia

O pré-processamento dos dados foi realizado por meio das ferramentas dispostas no pacote Bioconductor [2], disponíveis para a linguagem R. Essa etapa consiste em remover amostras de baixa qualidade, aplicar a normalização quantílica e remover sondas de sítios CpGs (regiões propensas a altas taxas de metilação) referentes ao sexo do paciente e/ou com reatividade cruzada. Vale apontar que o banco de dados apresenta os níveis de metilação $\beta = \frac{M}{M+U}$ de mais de 850 mil sítios CpGs, para cada um dos 407 pacientes [3], e portanto, para reduzir as dimensões dos dados em componentes, com o objetivo de facilitar a modelagem e a visualização dos mesmos, utilizou-se a técnica de Uniform Manifold Approximation and Projection (UMAP).

Na etapa de modelagem, utilizou-se diversas técnicas de aprendizado não-supervisionado [4], mais especificamente de agrupamentos, como K-Médias, Partition Around Medoids (PAM), Agrupamento Hierárquico e Misturas Finitas de Modelos.

No método K-Médias, buscar o melhor agrupamento é entendido como buscar pela partição $\mathcal{C}_1, ..., \mathcal{C}_K$ das observações, tal que se obtenha o menor valor possível para o somatório $\sum_{k=1}^{K} \frac{1}{|\mathcal{C}_k|} \sum_{i,j \in \mathcal{C}_k} d^2(x_i, x_j)$. Para escolhermos o melhor $K$, utiliza-se diversos métodos, como Método de Elbow e Silhueta.

O algoritmo PAM é muito similar ao K-Médias, o que o faz também ser conhecido como K-Medoids, pois o mesmo busca encontrar um *elemento central* dentro das próprias observações que minimize a distância entre as observações mais próximas, formando assim um grupo.

O método de Misturas Finitas de Modelos nos diz que, dado um número $G$ de variáveis aleatórias com distribuição $f_k(\mathbf{x})_i$, tal que $i = 1, 2, .., G$ e $\mathbf{x} = x_1, ..., x_n$ é uma amostra independente e identicamente distribuída, podemos escrever a distribuição de cada uma das observações por meio de uma função de densidade de probabilidade através de uma mistura de modelos de G componentes, isto é, $f(\mathbf{x}_i; \mathbf{\Psi}) = \sum_{k=1}^{G} \pi_k, = f_k(\mathbf{x}_i; \theta_k)$, onde $\mathbf{\Psi} = \pi_1, ..., \pi_{G-1}, \theta_1, ..., \theta_G$ são os parâmetros do modelo de misturas, $f_k(\mathbf{x}_i; \theta_k)$ é a k-ésima componente de densidade para a observação $i$ com parâmetro $\theta_k$, $(\pi_1, ..., \pi_{G-1})$ são os pesos ou probabilidades ($\pi_k > 0$, $\sum_{k=1}^{G} \pi_k = 1$) e $G$ é o número de componentes. Fixando-se G, podemos estimar os parâmetros $\mathbf{\Psi}$ maximizando a função do log da verossimilhança, dada por$\mathcal{L}(\mathbf{\Psi}; \mathbf{x}_1, ..., \mathbf{x}_n) = \sum_{i=1}^{n} log(\sum_{k=1}^{G} \pi_k f_k(\mathbf{x}_i; \theta_k))$.

## Resultados

Na figura 1, é possível ver que o pré-processamento uniformizou as densidades dos betas de cada um dos indivíduos.

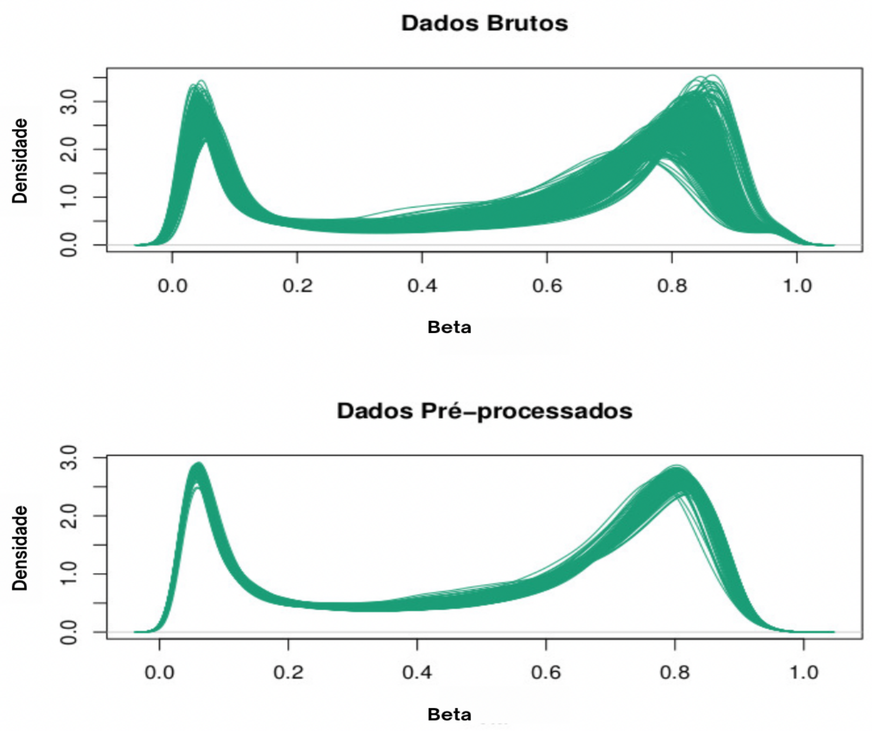


**Figura 1:** *Densidade das taxas de metilação antes e depois do pré-processamento.*

Além disso, as técnicas de aprendizado de máquina não-supervisionado junto ao UMAP permitiram identificar duas nuvens de pontos, expostas na figura 2.

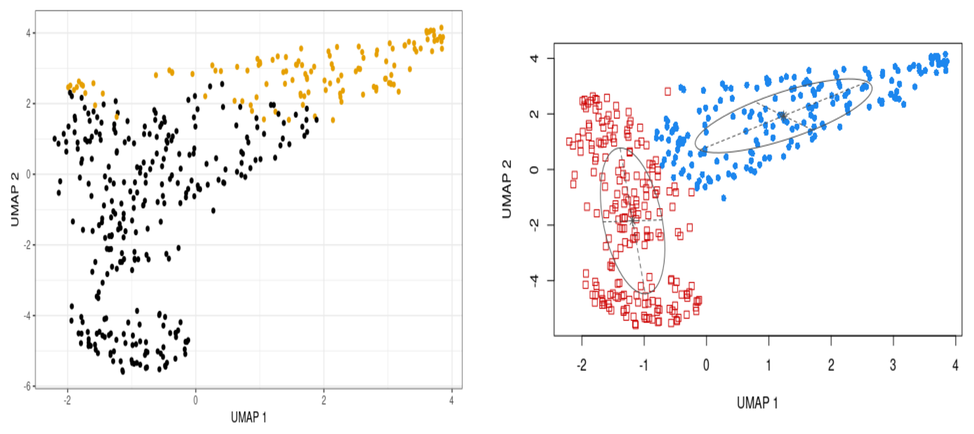


**Figura 2:** *K-Médias (à esquerda). Mistura de Modelos Gaussianos (EVV) (à direita).*

## Conclusão

De forma geral, podemos dizer que, utilizando-se de todos os sítios 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 agrupamento feito nas componentes do UMAP permitiu-nos encontrar 2 grupos significativos, cuja silhueta ficou próxima de 0,5. Devido à falta da indicação acerca da gravidade da COVID-19 no banco de dados utilizado, não podemos checar se os grupos formados podem ter sido decorrentes do grau dessa doença.

## Agradecimentos

Ao PIBIC/CNPq, pelo financiamento da pesquisa, e à professora Samara F. Kiihl, pela orientação.

**Referências**: [1] E R Gibney e C M Nolan. Epigenetics and gene expression. [2] W. Huber et al. Orchestrating high-throughput genomic analysis with Bioconductor. [3] Manuel Castro de Moura et al. Epigenome-wide association study of COVID-19 severity with respiratory failure. [4] R. Gentleman e V. J. Carey. Unsupervised Machine Learning.