Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística

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

Relatório Final

Aluno: Guilherme Pereira de Freitas
Orientadora: Samara Flamini Kiihl

Campinas
Setembro/2022

Sumário

1 Resumo

O presente trabalho tem como objetivo explorar as principais técnicas de pré-processamento de dados de metilação de DNA, bem como aplicar métodos de aprendizado de máquina não-supervisionado em dados de metilação de pacientes de Covid-19. Os pacientes estavam distribuídos em dois grupos: sintomas leves e severos. Dessa forma, o objetivo da pesquisa é verficar se o grau da comorbidade afeta a metilação do DNA dos pacientes, separando-os em dois grandes grupos. Além disso, passaremos por algumas técnicas computacionais que nos ajudarão a cumprir com o objetivo, como o gráfico de *elbow* junto à *silhueta*, para definir o melhor número de clusters.

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

2 Introdução

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 epigênicos. Neste projeto, iremos aplicar métodos de agrupamento em dados de metilação de DNA de pacientes de COVID-19, com o objetivo de encontrar grupos que podem ter sido formados por influência dessa doença.

O mecanismo estudado é a metilação da citosina (5mC), que ocorre 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 (Heyn e Esteller 2012). Em mamíferos, as citosinas metiladas estão restritas às CpGs (citosina-fosfato-guanina), onde as mesmas antecedem uma guanina (G) na direção de 5'. Vale lembrar que quatro bases nitrogenadas compõem a formação do DNA, 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 a outras regiões arbitrárias (Gibney e Nolan 2010).

3 Conceitos de Genética

Os primeiros conceitos de genética surgiram em meados do século XIX, através dos famosos experimentos com ervilhas conduzidos por Gregor Johann Mendel (1822 - 1884). Através de análises qualitativas e quantitativas em dados coletados durante uma década, Mendel foi capaz de mostrar que os indivíduos herdavam as características

de seus pais de modo previsível, além de concluir que cada particularidade de uma ervilha é controlada por um par de fatores que segrega-se na formação dos gametas (Klug et al. 2019).

A palavra "genética" foi introduzida somente em 1906, numa carta redigida pelo biólogo britânico Willian Bateson (1861-1926), com o objetivo de designar uma nova ciência de variação e hereditariedade. Baseada nos métodos probabilísticos de Mendel, essa nova ciência era distinguida pelo seu propósito explícito de generalizar a heredirariedade. Em 1910, a genética Mendeliana fundiu-se com a Teoria Cromossômica de Herança, dando início à conhecida Genética Clássica, que se dissolveria algumas décadas depois com a descoberta do DNA como base da herança genética. Esse último evento abriu as portas para a famosa Biologia Molecular, ciência moderna que utilizamos nos dias de hoje (Gayon 2016).

3.1 Células

Uma célula é formada por um conjunto de organelas que desempenham funções vitais para o seu funcionamento. Na imagem a seguir, é possível ver a estrutura celular de organismos eucariontes e procariontes, respectivamente (Clark, Pazdernik, e McGehee 2018).

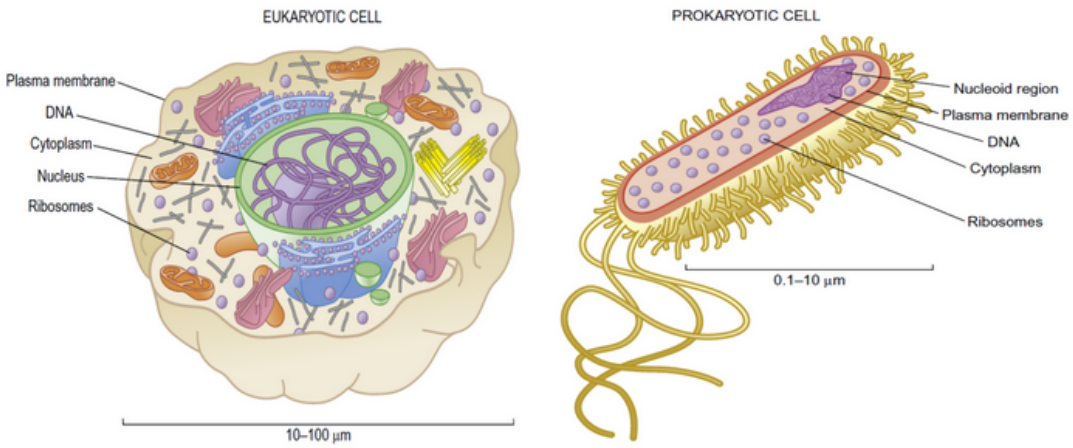

Figura 1: Célula eucarionte (animal) à esquerda e procarionte à direita.

Todas as células possuem características básicas. Elas são envolvidas pela membrana plasmática e apresentam uma substância gelatinosa, denominada citosol, que suspende os componentes celulares. No entanto, a maior diferença entre uma célula procariota e eucariota se dá na localização de seu DNA, pois em uma célula eucarionte, a maior parte do conteúdo genético está presente no núcleo, protegido por uma membrana, enquanto

que na procarionte, não temos um núcleo real ou uma membrana desempenhando a função de proteção (Reece 2020).

3.2 DNA

O DNA (Ácido Desoxirribonucleico) é a principal molécula portadora de informações dentro de uma célula, e sua estrutura se dá pela famosa dupla-hélice. Uma molécula de DNA de fita única, também chamada de polinucleotídeo, é uma cadeia de pequenas moléculas denominadas nucleotídeos. Cada nucleotídeo é formado a partir da combinação de três componentes; um açúcar de 5 carbonos (desoxirribose), um grupo fosfato e uma base nitrogenada.

Existem quatro bases nitrogenadas no DNA, que estão distribuídas em dois grupos; as purinas e as piridiminas. As purinas são as guaninas e adeninas, representadas respectivamente pelas letras G e A. O grupo das piridiminas é formado pelas citosinas e timinas, representadas respectivamente pelas letras C e T. Os polinucleotídeos podem ser formados por qualquer sequência de bases. Além disso, em uma dupla-hélice, as duas fitas de polinucleotídeos se ligam através da seguinte forma: C liga com G e A liga com T.

O fim de um polinucleotídeo é marcado por 5' ou 3'. Por convensão, uma fita de DNA é escrita com 5' no polo esquerdo e 3' no polo direito, de tal forma que dois polinucleotídeos são complementares se um pode ser obtido pelo outro através da troca mutual de A por T e C por G (Brazma et al. 2021).

3.3 RNA

Assim como o DNA, uma molécula de RNA (Ácido Ribonucleico) é formada por cadeias de nucleotídeos, mas utiliza a uracila (U) no lugar da timina (T) e o açúcar que a forma é a ribose. Essa diferença faz com que a mesma seja composta por um único polinucleotídeo, que leva a necessidade de uma estrutura mais complexa para realizar as ligações entre as bases (Brazma et al. 2021).

Existem diferentes moléculas de RNA, que desempenham funções importantes para a síntese de proteínas dentro de uma célula. Esse processo pode ser majoritariamente dividido em duas etapas; transcrição e tradução. De forma resumida, uma enzima, denominada RNA polimerase, inicia o processo de transcrição após reconhecer uma zona de interesse (o início de um gene, por exemplo). Após essa etapa, ela divide temporiariamente a dupla-hélice do DNA em dois polinucleotídeos e transcreve a sequência de um deles em uma molécula de RNA, que por sua vez é copiada em um RNA mensageiro (mRNA) (Tompa 2003).

Então, o mRNA se liga a um ribossomo e encontra o RNA transportador (tRNA), que reconhece as informações contidas no mRNA e carrega os aminoácidos apropriados

para a construção de proteínas durante a tradução (Klug et al. 2019).

3.4 Cromossomos e genes

A vida depende da capacidade das células de guardar, recuperar e traduzir as instruções genéticas para gerar e manter um organismo vivo (Alberts et al. 2002). Essas instruções são carregadas por moléculas de DNA, que por sua vez formam as estruturas complexas denonimadas cromossomos. Nos cromossomos, o DNA se apresenta enrolado em proteínas chamadas de histonas, de tal forma que se fosse esticado alcançaria 1m de comprimento. Além disso, dentro de um organismo multicelular, todas as células carregam o mesmo conteúdo genético, com algumas poucas excessões, devido ao resultado da replicação de DNA em cada divisão celular (Brazma et al. 2021).

Todos os seres humanos apresentam 23 pares de cromossomos, e a diferenciabilidade entre as suas células, como em qualquer organismo multicelular, ocorre através da regulação da expressão gênica, que pode silenciar determinados genes para obter tipos específicos de células (como a da pele, por exemplo).

Existem várias definições do significado de um "gene". Brazma et al. aponta que um gene é um trecho contínuo de uma molécula de DNA, a partir do qual um complexo maquinário molecular pode ler informações (codificadas com as letras A, T, G e C) e fazer um tipo específico ou um conjunto de proteínas. Os genes são fundamentais em vários processos biomoleculares, como na síntese protéica, onde RNA polimerase identifica a sequência de bases de um gene específico e inicia o processo de produção das proteínas requisitadas.

4 Epigenética e Expressão Gênica

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 (Gibney e Nolan 2010).

O mecanismo estudado é a metilação da citosina (5mC), que ocorre em áreas específicas de regulação gênica, como regiões promotoras ou de heterocromatina. Níveis anômalos de metilação podem, assim, inibir vários genes supressores de tumor, já que diferentes tipos e subtipos de tumores apresentam perfis distintos de metilação do DNA nas ilhas CpG (Damgacioglu, Celik, e Celik 2019). Em mamíferos, as citosinas metiladas estão restritas às CpGs, 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 quando comparadas com outras regiões arbitrárias (Gibney e Nolan 2010). A Figura 2 retrata a adição do grupo metil CH_3 à estrutura química de uma citosina (Saini et al. 2013).

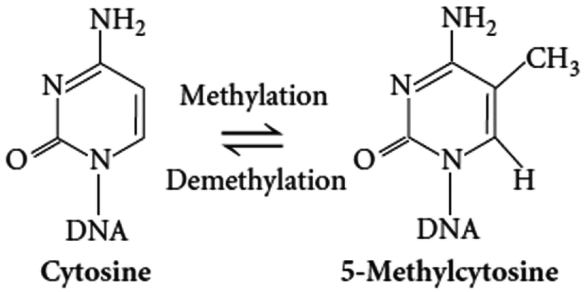

Figura 2: Metilação da citosina para 5-metilcitosina, que impede sua transcrição para uracila

5 Microarranjos

Microarranjos de DNA são arranjos de estruturas fixas de ácido nucleico, chamadas de sondas, cujos padrões foram definidos durante a construção ou depositados em um substrato sólido e plano, geralmente de vidro ou silício. Essas plataformas são utilizadas para investigar a quantidade de mRNA, ou genes expressos, presente na amostra biológica sob o experimento (experimento de hibridização). Atualmente, existe uma tendência em usar o sequênciamento de genes com o objetivo de desenvolver sondas e possibilitar a fabricação de microarranjos (Scherer 2009).

6 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 (Moura et al. 2021). 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, com idade máxima de 61 anos, onde 194 (47.7%) foram diagnosticados com sintomas leves e 213 (52.3%) com sintomas severos. Os dados foram coletados através da plataforma *Infinium MethylationEPIC BeadChip*, totalizando em 850 mil ilhas CpGs para cada indivíduo. É importante mencionar que não há indicação dos sintomas dos pacientes nos dados disponibilizados pelos pesquisadores.

6.1 Infinium MethylationEPIC BeadChip

O Infinium MethylationEPIC BeadChip, exposto na Figura 3, é o novo chip da Illumina, sucessor do Illumina HumanMethylation450 (HM450) BeadChip, que cobria aproximadamente 450.000 CpGs. Esse dispositivo cobre cerca de 90% das CpGs de HM450 e um adicional de 413.743, somando mais de 850 mil ilhas. Isso é possível devido ao uso das sondas Infinium II, que necessitam apenas de uma sonda por sítio. Além disso, das 413.743 CpGs adicionais, 95% utilizam as novas sondas. A alta proporção de sondas do tipo II ocupa menos espaço, maximizando sua quantidade, porém reduz o número de amostras mensuradas pelo chip de 12 (HM450) para 8 (EPIC) (Pidsley et al. 2016).

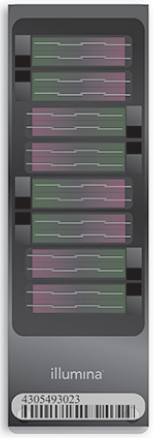

Figura 3: O Infinium MethylationEPIC BeadChip apresenta > 850.000 CpGs em regiões potenciadoras, corpos gênicos, promotores e ilhas CpG. (Illumina 2015)

Para cada sítio CpG, o chip registra suas intensidades de metilado e não metilado, de modo que os níveis de metilação são obtidos através da fórmula $\beta = \frac{M}{M+U}$, tal que M é a intensidade de metilado e U é a intensidade de não metilado. Outra métrica muito utilizada para medir o nível de metilação é dada por $Mvalue = log2(\frac{M}{U})$. É muito comum somar um α ao denominador de β, para evitar cenários de divisão por zero quando $M + U \longrightarrow 0$ e arrumar a escala dos coeficientes (Maksimovic, Phipson, e Oshlack 2016).

7 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" (Maksimovic, Phipson, e Oshlack 2016), por meio das ferramentas oferecidas no pacote Bioconductor (Huber et al. 2015), disponíveis para a linguagem R (R Core Team 2020). Os agoritmos são aplicados na matriz de intensidades (betas), cujo cálculo será descrito na seção seguinte. Vale enfatizar que o controle de qualidade das amostras é vital para a análise dos dados, pois permite minimizar enviesamentos e ter mais confiança em realizar alguma conclusão sobre o efeito da Covid-19 nos pacientes.

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 (Freitas 2022).

7.1 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 μ é a mediana e σ^2 é o desvio absoluto mediano das intensidades para essas sondas. Após isso, devemos obter a intensidade de metilação total do indivíduo em cada sítio e calcular a probabilidade de cada uma dessas intensidades ser uma amostra da distribuição normal obtida no início (Aryee et al. 2014).

7.2 Filtro das amostras

O primeiro filtro de qualidade é aplicado com o intuito de remover as amostras de baixa qualidade. Para cada indivíduo, vemos se a média dos p-valores é menor que um nível de significância α. Aqui, adotamos $\alpha = 0.05$, por recomendação do artigo de referência.

7.3 Normalização quantílica

A normalização quantílica (Touleimat e Tost 2012) é uma técnica de pré-processamento que realiza diversas correções no conjunto de dados. Sua pipeline é composta, respectivamente, pelas etapas de controle de qualidade, filtro das sondas, correção de

sinais e normalização quantílica baseada em subconjuntos. A etapa de controle de qualidade estuda os efeitos de laboratório para estimar a qualidade das sondas e das amostras, já a etapa de filtro consiste em remover as sondas cuja variação do nível de metilação pode ocorrer devido a variações genéticas. A etapa de correção de sinais aplica uma normalização quantílica suave para corrigir possíveis problemas de marcação e escaneamento dos canais de cores. Por fim, a última etapa aplica uma normalização robusta para corrigir possíveis enviesamentos, nos valores de betas, causados pelo uso dos dois tipos de ensaios (Inf I e Inf II) no chip do experimento.

7.4 Filtro das sondas

Nessa etapa, aplicou-se diversos filtros diferentes. O primeiro é mais simples, e calcula a média dos p-valores dos indivíduos, fixando-se ilha por ilha, e segue apenas com as CPG's que registrarem valores inferiores a $\alpha = 0.01$. O segundo filtro tem como objetivo remover as sondas dos cromossomos X e Y, para evitar possíveis tendências de metilação dadas pelo sexo do paciente.

O terceiro filtro busca remover as sondas afetadas por SNPs (Single Nucleotide Polymorphism) em seus campos, com o objetivo de evitar possíveis enviesamentos, pois o nível de metilação captado pelo sinal pode ser decorrente de CpGs polimórficas que sobrepuseram regiões de SNPs. Por último, é importante remover as sondas que demonstraram ser reativo-cruzadas, ou em inglês, *cross-reactive*, pois as mesmas se ligam a múltiplos trechos do genoma (Chen et al. 2013).

7.5 Resultado final

Na Figura 4, é possível ver que o pré-processamento uniformizou as densidades dos betas de cada um dos indivíduos. Portanto, temos mais confiança para dizer que os efeitos de laboratório e da coleta de dados foram reduzidos e apresentam menos impacto em nossas análises. Em suma, as etapas de pré-processamento consistiram em remover amostras de baixa qualidade, aplicar a normalização quantílica e remover sondas de CpGs referentes ao sexo do paciente e/ou com reatividade cruzada.

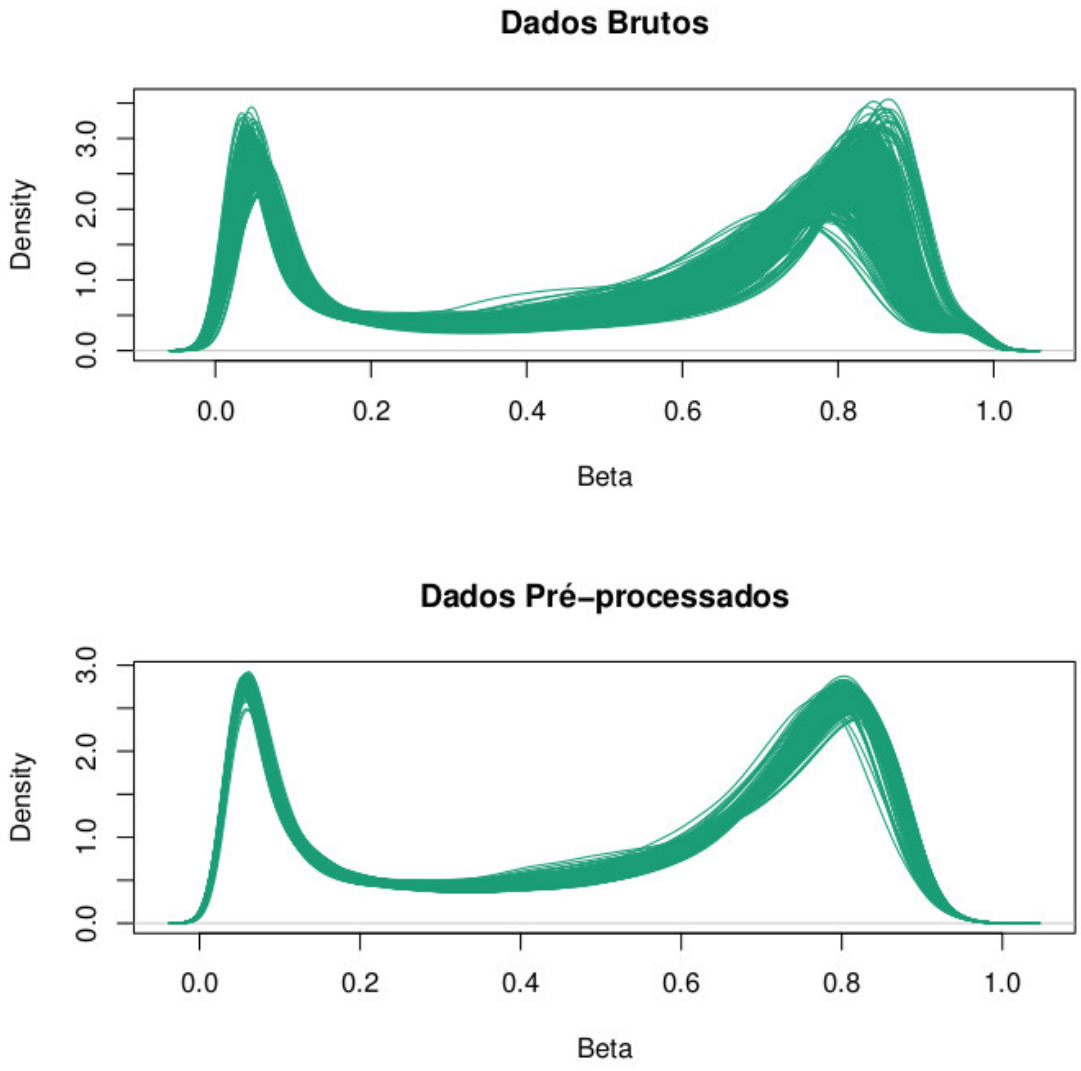

Figura 4: Densidade das taxas de metilação antes e depois do pré-proprocessamento.

8 Aprendizado de Máquina Não-supervisionado

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 um *target* realizar o treinamento. Os pré-requisitos para aplicar as técnicas de agrupamento se dão na escolha das variáveis, hiperparâmetros e tipo de distância adotada (Gentleman e Carey 2008).

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).

Para avaliar o número de grupos, utilizou-se as técnicas de Silhueta, Gráfico de Elbow e BIC.

- Silhueta: Dado um conjunto de clusteres Λ, temos que a silhueta da observação i presente no cluster λ_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 λ_k (que o contém) e b_i é a menor dissimilaridade de i com relação aos elementos de outro cluster λ, ou seja, $b_i = min_{\lambda \neq \lambda_k} d(i, \lambda)$ (Rousseeuw 1986).
- Método de Elbow: O Método de Elbow (Hayasaka 2022) é uma curva construída a partir da Soma de Quadrados Intra-cluster, ou inércia, cuja fórmula é dada por $WSS = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$ (Imran 2015). O melhor número de clusteres é obtido no ponto de maior inclinação da curva.
- Critérios de Informação: Para avaliar o número de grupos no método de mistura de modelos, utilizou-se o Critério de Informação Bayesiana (BIC) (Vrieze 2012).

8.1 K-Médias

O método das K-Médias (Lloyd 1982) é uma técnica de agrupamento que funciona a partir de um parâmetro inicial, o número de clusteres. Nesse algoritmo, 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 $\sum_{k=1}^{K} \frac{1}{|\mathcal{C}_k|} \sum_{i,j \in \mathcal{C}_k} d^2(x_i, x_j)$ (Rafael Izbicki 2020).

Esse somatório representa a soma de quadrados dentro de cada cluster. O algoritmo de K-Médias é iterativo e funciona da seguinte forma: na primeira iteração, toma-se k amostras (centroids) e liga cada uma das observaçõs ao centroid mais próximo; em seguida calcula-se a média das observações de cada cluster e obtém k novos centroids. Esses passos são executados até minimizar a soma anterior.

8.2 PAM

O algoritmo PAM - Partition Around Medoids (Kaufman Rousseeuw 2009) 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, chamado de medoid, que minimiza a distância entre as observações mais próximas, formando assim um cluster. Para escolher os novos medoids em cada cluster, obtêm-se o ponto que produz a menor soma de dissimilaridade com relação às outras observações. Com os novos medoids determinados, aloca-se todas as observações ao cluster mais próximo e repete as etapas anteriores.

O medoid j, por exemplo, é dado por $c_j \in (x_1, x_2, ..., x_{n-1}, x_n)$, onde x_i é a observação i. Os passos do método pam podem ser descritos da seguinte forma:

1. Escolha uma amostra de tamanho k (medoids) dentro das observações.
2. Para cada observação da base de dados, relacione-a com o medoid mais próximo.
3. Para cada cluster formado pelo medoid j, ache o elemento que reproduz a menor dissimilaridade entre os outros do mesmo grupo, e transforme-o no novo medoid.
4. Reproduza as etapas 2 e 3 até que os medoids não mudem de uma iteração para outra.

8.3 Clustering Hierárquico

Os métodos de clustering hierárquico funcionam de tal forma que, dado um cojunto de g clusters, ao obtermos outro cojunto de g+1 a partir do inicial temos que ambos conjuntos de clusters apresentam g-1 grupos idênticos, e o grupo remanecente é dividido em dois. Existem diversas técnicas dessa família; as mais famosas são a de Single-Linkage, que considera a distância entre dois grupos como sendo a distância entre seus pontos mais próximos, e a Complete-Linkage, que considera a distância entre dois grupos como sendo a distância entre seus dois pontos mais distantes. Ambas as técnicas apresentam diferentes versões de agrupamentos, como a aglomerativa e a divisa (Mardia, Bibby, e Kent 1979).

8.4 Modelos de Misturas Gaussianas

Misturas finitas de modelos (Scrucca et al. 2016) estão cada vez mais sendo utilizadas em diversos ramos de pesquisa, como análise de agrumaneto, classificação e estimação de densidades. O método nos diz que, dado um número G de variáveis aleatórias com distribuição $f_k(\mathbf{x})_i$, tal que $i = 1, 2, 3, .., G$ e $\mathbf{x}$=$\{x_1, x_2, ..., x_n\}$ 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 finita de modelos de G componentes da seguinte forma

$$f(\mathbf{x}_i; \mathbf{\Psi}) = \sum_{k=1}^{G} \pi_k f_k(\mathbf{x}_i; \theta_k)$$

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 θ_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 da mistura.

Assumindo que G é pré-fixado, os parâmetros $\mathbf{\Psi}$ da mistura de modelos são, geralmente, desconhecidos e precisam ser estimados. A função do log da verossimilhança é dada por

$$\mathcal{L}(\mathbf{\Psi}; \mathbf{x}_1, ..., \mathbf{x}_n) = \sum_{i=1}^{n} log(f(\mathbf{x}_i; \mathbf{\Psi})) = \sum_{i=1}^{n} log(\sum_{k=1}^{G} \pi_k f_k(\mathbf{x}_i; \theta_k)).$$

A maximização direta da verossimilhança é muito complicada devido ao logaritmo da soma. Para contornar esse problema, utiliza-se a técnica *Expextation - Maximization*, ou simplesmente EM (Borman 2004).

No presente trabalho, utilizou-se um modelo popuar denominado Gaussian Mixture Model (GMM), que assume uma distribuição gaussiana multivariada para cada componente, isto é, $f_k(\mathbf{x}; \theta_k) \sim \mathcal{N}(\mu_k, \mathbf{\Sigma}_k)$. Os agrupamentos formados são elipsoidais, centrados no vetor de médias μ_k, com propriedades geométricas definidas pela matriz de covariâncias $\mathbf{\Sigma}_k$. O pacote mclust, utilizado na pesquisa, inicia o algoritmo aplicando-se um clustering hierárquico, para obter Ψ_o, e então atualiza sequencialmente as densidades buscando-se maximizar a função $\mathcal{L}(\mathbf{\Psi}; \mathbf{x}_1, ..., \mathbf{x}_n)$.

8.5 Redução de Dimensões

As técnicas de redução de dimensão estão sendo aplicadas em um variedade de campos e em tamanhos cada vez maiores de conjuntos de dados. UMAP (Uniform Manifold Approximation and Projection) (McInnes, Healy, e Melville 2020) é uma técnica de aprendizado múltiplo para redução de dimensões. UMAP foi construído a partir de um framework baseado na geometria de Riemannian e topologia algébrica, e fornece um algoritmo escalável e aplicável em dados do mundo real. Além disso, essa técnica é uma alternativa ao t-SNE, fornecendo mais qualidade de visualização, preservando mais da estrutura global dos dados e sendo mais eficiente no tempo de execução.

9 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. Em cada um dos gráficos, os clusteres (grupos) encontrados são representados por diferentes cores, sendo amarelo e preto nas duas primeiras figuras e vermelho e azul na última.

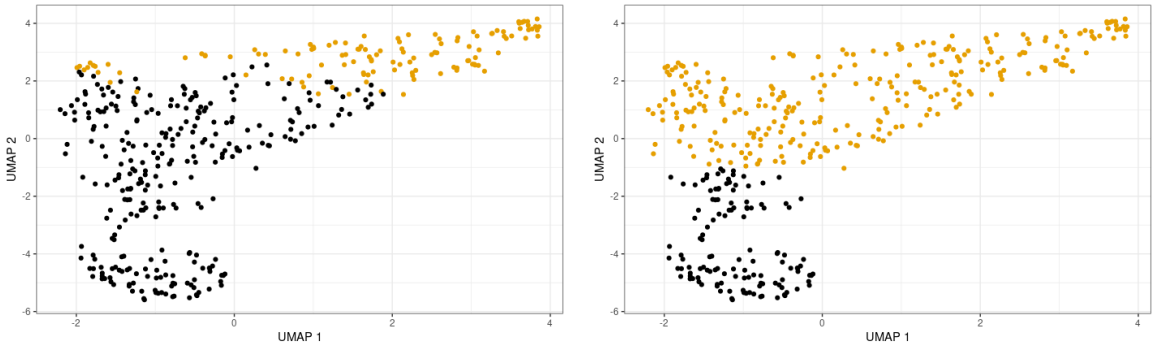

Figura 5: K-Médias: Agrupamentos encontrados através da matriz de distâncias completa (à esquerda) e através da matriz de distâncias das componentes UMAP (à direita). Em ambos os gráficos foram utilizadas as componentes do UMAP para gerar as visualizações.

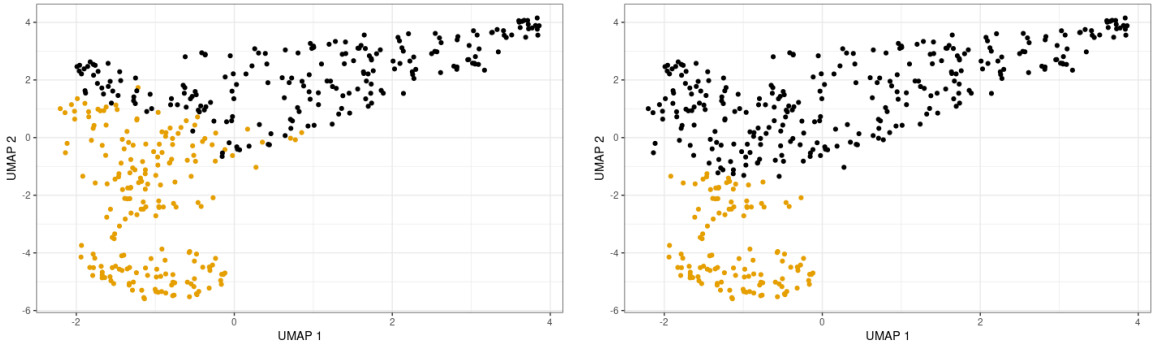

Figura 6: PAM: Agrupamentos encontrados através da matriz de distâncias completa (à esquerda) e através da matriz de distâncias das componentes UMAP (à direita). Em ambos os gráficos foram utilizadas as componentes do UMAP para gerar as visualizações.

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).

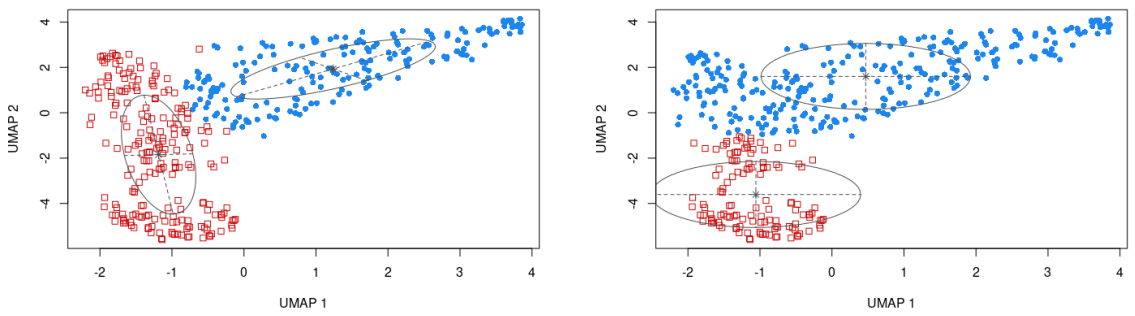

Figura 7: Mistura de Modelos Gaussianos: Agrupamentos encontrados através da matriz de distâncias das componentes UMAP. EVV (à esquerda) e EII (à direita). Em ambos os gráficos foram utilizadas as componentes do UMAP para gerar as visualizações.

10 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 da indicação acerca da gravidade da COVID-19 no banco de dados utilizado, pois essa informação não foi disponibilizada pelos autores da pesquisa, não podemos checar se os grupos formados podem ter sido decorrentes do grau dessa doença.

Referências

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, e Peter Walter. 2002. *Molecular Biology of the Cell, Fourth Edition*. 4º ed. Garland Science.

Aryee, Martin J., Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, e Rafael A. Irizarry. 2014. "Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays". *Bioinformatics* 30 (10): 1363–69. https://doi.org/10.1093/bioinformatics/btu049.

Borman, Sean. 2004. "The expectation maximization algorithm-a short tutorial". *Submitted for publication* 41.

Brazma, Alvis, Helen Parkinson, Thomas Schlitt, e Mohammadreza Shojatalab. 2021. "A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays", outubro.

Chen, Yi-an, Mathieu Lemire, Sanaa Choufani, Darci T. Butcher, Daria Grafodatskaya, Brent W. Zanke, Steven Gallinger, Thomas J. Hudson, e Rosanna Weksberg. 2013. "Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray". *Epigenetics* 8 (2): 203–9. https://doi.org/10.4161/epi.23470.

Clark, David P., Nanette Pazdernik, e Michelle McGehee. 2018. *Molecular Biology*.

Damgacioglu, Haluk, Emrah Celik, e Nurcin Celik. 2019. "Estimating gene expression from high-dimensional DNA methylation levels in cancer data: A bimodal unsupervised dimension reduction algorithm". *Computers & Industrial Engineering* 130: 348–57. https://doi.org/https://doi.org/10.1016/j.cie.2019.02.038.

Freitas, Guilherme. 2022. "Projeto de Iniciação Científica". *Github*. https://github.com/GuilhermeFreitas09/IC.

Gayon, Jean. 2016. "From Mendel to epigenetics: History of genetics". *Comptes Rendus Biologies* 339 (junho). https://doi.org/10.1016/j.crvi.2016.05.009.

Gentleman, R., e V. J. Carey. 2008. "Unsupervised Machine Learning". In *Bioconductor Case Studies*, 137–57. Springer New York. https://doi.org/10.1007/978-0-387-77240-0_10.

Gibney, E R, e C M Nolan. 2010. "Epigenetics and gene expression". *Heredity* 105 (1): 4–13. https://doi.org/10.1038/hdy.2010.54.

Hayasaka, Satoru. 2022. "How many clusters?". *Medium*. Towards Data Science. https://towardsdatascience.com/how-many-clusters-6b3f220f0ef5.

Heyn, Holger, e Manel Esteller. 2012. "DNA methylation profiling in the clinic: applications and challenges". *Nature Reviews Genetics* 13 (10): 679–92. https://doi.org/10.1038/nrg3270.

Huber, W., V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, et al. 2015. "Orchestrating high-throughput genomic analysis with Bioconductor". *Nature Methods* 12 (2): 115–21. http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html.

Imran. 2015. "What is "within cluster sum of squares by cluster" in k-means". *Data Science, Analytics and Big Data discussions*. https://discuss.analyticsvidhya.com/t/what-is-within-cluster-sum-of-squares-by-cluster-in-k-means/2706.

Kaufman Rousseeuw, Leonard Peter J -. 2009. *Finding Groups in Data*. John Wiley & Sons, Inc.

Klug, William, Michael Cummings, Charlotte Spencer, Michael Palladino, e Darrell Killian. 2019. *Concepts of Genetics (Masteringgenetics)*. 12º ed. Pearson.

Lloyd, S. 1982. "Least squares quantization in PCM". *IEEE Transactions on Information Theory* 28 (2): 129–37. https://doi.org/10.1109/TIT.1982.1056489.

Maksimovic, Jovana, Belinda Phipson, e Alicia Oshlack. 2016. "A cross-package Bioconductor workflow for analysing methylation array data". *F1000Research* 5 (junho): 1281. https://doi.org/10.12688/f1000research.8839.1.

Mardia, K V, J M Bibby, e J T Kent. 1979. *Multivariate analysis*. Livro. http://www.loc.gov/catdir/toc/els031/79040922.html.

McInnes, Leland, John Healy, e James Melville. 2020. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". http://arxiv.org/abs/1802.03426.

Moura, Manuel Castro de, Veronica Davalos, Laura Planas-Serra, Damiana Alvarez-Errico, Carles Arribas, Montserrat Ruiz, Sergio Aguilera-Albesa, et al. 2021. "Epigenome-wide association study of COVID-19 severity with respiratory failure". *EBioMedicine* 66 (abril): 103339. https://doi.org/10.1016/j.ebiom.2021.103339.

Pidsley, Ruth, Elena Zotenko, Timothy J. Peters, Mitchell G. Lawrence, Gail P. Risbridger, Peter Molloy, Susan Van Djik, Beverly Muhlhausler, Clare Stirzaker, e Susan J. Clark. 2016. "Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling". *Genome Biology* 17 (1). https://doi.org/10.1186/s13059-016-1066-1.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rafael Izbicki, Tiago Mendonça Dos Santos. 2020. *Aprendizado de máquina: uma abordagem estatística*. Garland Science.

Reece, Jane Taylor. 2020. *Campbell Biology*. 12º ed. Pearson.

Rousseeuw, Peter J. 1986. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis".

Saini, Amarjit, Sarabjit Mastana, Fiona Myers, e Mark Lewis. 2013. "'From Death, Lead Me to Immortality' – Mantra of Ageing Skeletal Muscle". *Current genomics* 14 (junho): 256–67. https://doi.org/10.2174/1389202911314040004.

Scherer, Andreas. 2009. *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. 1º ed. Wiley.

Scrucca, L., M. Fop, T. B. Murphy, e A. E. Raftery. 2016. "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models". *R J* 8 (1): 289–317.

Tompa, Martin. 2003. "Basics of Molecular Biology", julho.

Touleimat, Nizar, e Jörg Tost. 2012. "Complete pipeline for Infinium Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation". *Epigenomics* 4 (3): 325–41. https://doi.org/10.2217/epi.12.21.

Vrieze, Scott I. 2012. "Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)". *Psychol. Methods* 17 (2): 228–43.