

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

Relatório Parcial

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

28/02/2022

1 Resumo

O presente trabalho tem como objetivo explorar as principais técnicas de pré-processamento de dados de metilação de DNA, bem como utilizar e desenvolver métodos de Aprendizado de Máquina não Supervisionado em dados de metilação de pacientes de Covid-19 (Moura et al. 2021). Além disso, passaremos por algumas técnicas computacionais que nos ajudarão a cumprir com o objetivo, como o UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction), para redução de dimensões, e o gráfico de Elbow junto a Silhueta, para definir o melhor número de clusters. Todos os códigos desenvolvidos estão expostos no repositório do Github. \par Palavras-chave: bioinformática, metilação de DNA, aprendizado de máquina não supervisionado, métodos de agrupamento, COVID-19, EWAS

2 Introdução

Embora grande parte das células de um organismo multicelular apresente 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 será 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 (Illumina 2017). Em mamíferos, as citosinas metiladas estão restritas às CpGs (cytosine-phosphate-guanine), 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 (Gibney e Nolan 2010).

3 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 sequenciamento de genes com o objetivo de desenvolver sondas e possibilitar a fabricação de microarranjos. (Scherer 2009)

3.1 Infinium MethylationEPIC BeadChip

O Infinium MethylationEPIC BeadChip é o novo chip da Illumina, sucessor do Illumina HumanMethylation450 (HM450) BeadChip, que cobria aproximadamente 450.000 CpGs. O novo chip cobre mais 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 necessita apenas de 2 sondas (beads) por Locus. 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)

Para cada ilha CpG, o chip registra suas intensidades de metilado e não metilado, de modo que os níveis de metilação são obtidos a partir da seguinte forma:

$$\beta = \frac{M}{M + U}$$

Ta que M é a proporção de metilado e U é a proporção de não metilado. Outra técnica muito utilizada para medir o nível de metilação é dada por $Mvalue = \log_2(\frac{M}{U})$. É muito comum somar um α ao denominador de β , para evitar cenários de divisão por zero quando $M + U \rightarrow 0$.

4 Pré-processamento

O fluxo de pré-processamento será feito seguindo o passo a passo descrito no artigo “A cross-package Bioconductor workflow for analysing methylation array data” (Maksimovic, Phipson, e Oshlack 2016), por meio das ferramentas dispostas no pacote Bioconductor (Huber et al. 2015), disponíveis para a linguagem R (R Core Team 2020). Os algoritmos são aplicados na matriz de p-valores, 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 alguma conclusão sobre o efeito da Covid-19 nos pacientes.

4.1 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 (combinação de canais de cores) 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 ilha e calcular a probabilidade de cada uma dessas intensidades ser uma amostra da distribuição normal obtida no início.

4.2 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.

4.3 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.

4.4 Filtro das sondas

Nessa etapa, aplicou-se diversos filtros diferentes. O primeiro é mais simples, e cacula 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, para 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)

5 Aprendizado de Máquina não Supervisionado

Falar sobre o que é aprendizado não supervisionado...

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.

Para escolher o número de grupos, em cada um dos métodos, utilizaremos as técnicas de Silhueta e Gráfico de Elbow.

- Silhueta: Dado um conjunto de clusters 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). Temos evidência de formação de clusters quando a média desse score é superior a 0.4.
- Gráfico de Elbow: O Gráfico de Elbow é uma curva construída a partir da Soma de Quadrados Intra-cluster, cuja fórmula é dada por $WSS = \sum_{k=1}^K \sum_{i \in S_k} \sum_{j=1}^p (x_{ij} - \bar{x}_{kj})^2$ ("What is "within cluster sum of squares by cluster" in k-means" 2015). O melhor número de clusters é obtido no ponto de maior inclinação da curva.

5.1 KMEANS

2.1) Melhor número de clusters via elbow e silhueta

2.2) Modelagem

5.2 PAM

3.1) Melhor número de clusters via elbow e silhueta

3.2) Modelagem

- 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>.
- Gibney, E R, e C M Nolan. 2010. “Epigenetics and gene expression” 105 (1): 4–13. <https://doi.org/10.1038/hdy.2010.54>.
- 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>.
- Illumina. 2017. “An introduction to Next-Generation Sequencing Technology”, 16.
- 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>.
- 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 Dijk, 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/>.
- Rousseeuw, Peter J. 1986. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”.
- Scherer, Andreas. 2009. *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. 1^o ed. Wiley.
- 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>.
- “What is "within cluster sum of squares by cluster" in k-means”. 2015. *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.