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

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.

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

1 Introdução

Embora boa parte das células de um organismo multicelular apresentem 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, 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). A figura 1 retrata a adição do grupo metil CH_3 à estrutura química de uma citosina (Saini et al. 2013).

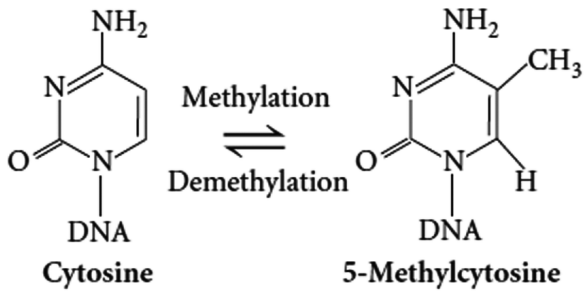

Figura 1: Metilação da citosina para 5-metilcitosina, que impede sua transcrição para uracila

2 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)

2.1 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 chipe 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 $Mvalue = log2(\frac{M}{U})$. É muito comum somar um α ao denominador de β, para evitar cenários de divisão por zero quando $M + U \longrightarrow 0$.

3 Pré-processamento

O fluxo de pré-processamento será 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 dispostas no pacote Bioconductor (Huber et al. 2015), disponíveis para a linguagem R (R Core Team 2020). Os agoritmos são aplicados na matriz de p-valores, cujo cálculo será descrito na seção seguinte.

3.1 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 de dados. O primeiro passo é filtrar as sondas de controle separadas para o paciente em questão, 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 probabiidade de cada uma dessas intensidades ser uma amostra da distribuição normal obtida no início.

3.2 Filtro das amostras

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

3.3 Normalização quantílica

A normalização quantílica realizada na presente pesquisa é capaz de...

3.4 Filtro das ilhas CpG's

Nessa etapa, aplicou-se dois 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 registraram médias inferiores a $\alpha = 0.01$.

- relativas a ilhas CpGs de sexo

3.5 Remoção das sondas com SNP nos campos de CpGs

3.6 Exclusão de sondas com reatividade cruzada

4 Aprendizado de Máquina não Supervisionado

- Matriz de distâncias
- Elbow
- Silhueta

4.1 KMEANS

```
2.1) Melhor número de clusters via elbow e silhueta
2.2) Modelagem
2.3) print do resultado usando as dimensões reduzidas
```

4.2 PAM

```
3.1) Melhor número de clusters via elbow e silhueta
3.2) Modelagem
3.3) print do resultado usando as dimensões reduzidas
```

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 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/.

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.