

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 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 ocorrem a partir 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 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).

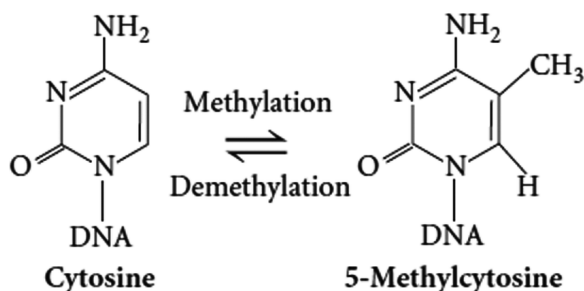


Figura 1: Metilação da citosina para 5-metilcitosina, que impede sua transcrição para uracila

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

2.1 Infinium MethylationEPIC BeadChip

Infinium MethylationEPIC BeadChip (Figura 2) 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)

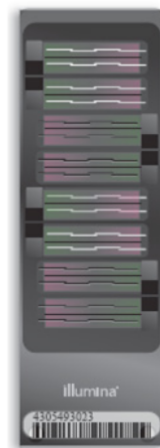


Figura 2: O Infinium MethylationEPIC BeadChip apresenta > 850.000 CpGs em regiões potenciadoras, corpos gênicos, promotores e ilhas CpG. (Illumina 2015)

3 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).

3.1 Filtro das amostras

3.2 Normalização quantílica

3.3 Filtro das sondas

- defeituosas
- relativas a ilhas CpGs de sexo

3.4 Remoção das sondas com SNP nos campos de CpGs

3.5 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.