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

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 epigenônicos. Neste projeto, iremos aplicar métodos de agrupamento em dados de metilação de DNA de pacientes de COVID-19. Palavras-chave: bioinformática, metilação de DNA, aprendizado de máquina não supervisionado, métodos de agrupamento, COVID-19, EWAS

1 Introdução

Nada por enquanto

2 Genética

Os primeiros conceitos de genética surgiram em meados do século XIX, através dos famosos experimentos em ervilhas conduzidos pelo biólogo Gregor Johann Mendel (1822 - 1884). Por meio 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)

No entanto, 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 era 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)

2.1 Introdução à Biologia Molecular e Mecanismos Genéticos

2.1.1 Células

Uma célula é formada por um conjunto de organelas que desempenham funções vitais para o seu funcioamento. Na imagem a seguir, é possível ver a estrutura celular de organismos eucariontes e procariontes, respectivamente.

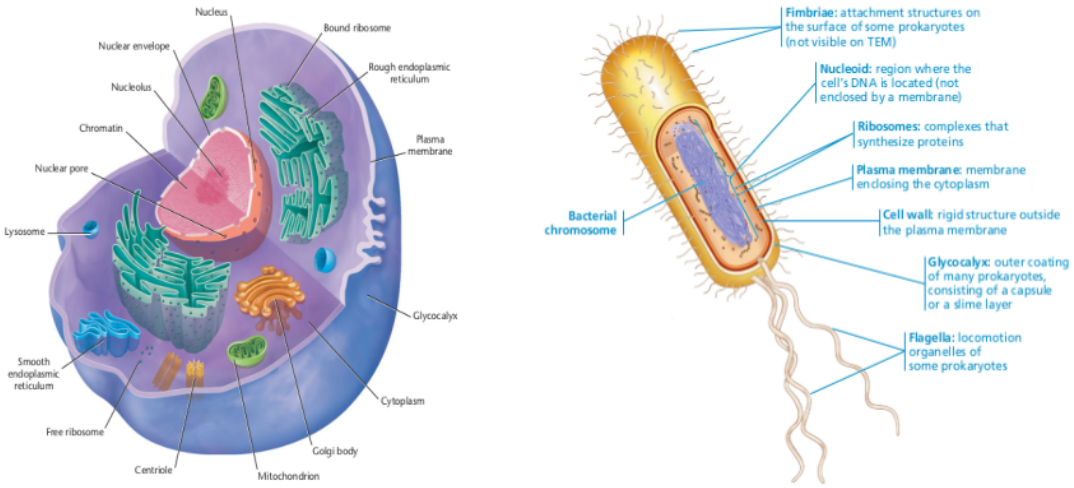

Figura 1: A figura à esquerda é uma célula eucarionte animal (Klug et al. 2019). A figura à direita representa uma célula eucarionte (Reece 2020)

Todas as células possuem características básicas. Elas são elvolvidas 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 com a função de proteção. (Reece 2020)

2.1.2 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. É comum utilizar o termo "base nitrogenada" para se referenciar a um nucleotídeo.

Existem 4 diferentes nucleotídeos, 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 formato pelas citosinas e timinas, representadas respectivamente pelas letras C e T. Os polinucleotídeos podem ser formados por qualquer sequência de bases e podem assumir tamanhos diversos. 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)

2.1.3 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). Essa diferença faz com que a mesma seja formada 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 será copiada em um RNA mensageiro (mRNA). (Tompa 2003)

Então, o mRNA se liga a um ribossomo e encontra o RNA transportador (tRNA), que por sua vez reconhece as informações contidas no mRNA e carrega os aminoácidos apropriados para a construção de proteínas durante a tradução. (Klug et al. 2019)

2.1.4 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)

Existem várias definições do significado de um "gene". (Brazma et al. 2021) 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.

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).

3 Epigenética e expressão gênica

Embora todas as 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 2 retrata a adição do grupo metil CH_3 à estrutura química de uma citosina (Saini et al. 2013).

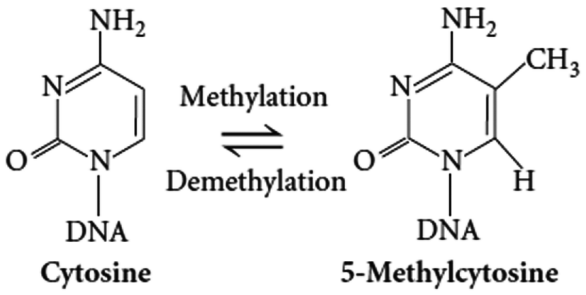

Figura 2: Metilação da citosina para 5-metilcitosina, que impede sua transcrição para uracila

4 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)

4.1 Infinium MethylationEPIC BeadChip

Infinium MethylationEPIC BeadChip (Figura 3) é o novo chip da Illumina, sucessor do Illumina HumanMethylation450 (HM450) BeadChip, que cobria aproximadamente 450.000 CpGs e utilizava a estrutura de sondas Infinium I. A nova tecnologia cobre mais de 90% das CpGs de HM450 e um adicional de 413.743, somando mais de 850 mil ilhas. Isso é possível por meio do uso das sondas Infinium II, que necessita apenas de 2 sondas por Locus. Além disso, das 413.743 CpGs adicionais, 95% utiliza 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)

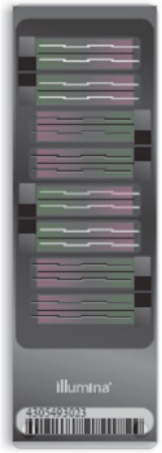

Figura 3: O Infinium MethylationEPIC BeadChip apresenta > 850.000 CpGs em regiões potenciadoras, corpos gênicos, promotores e ilhas CpG. (Illumina 2015)

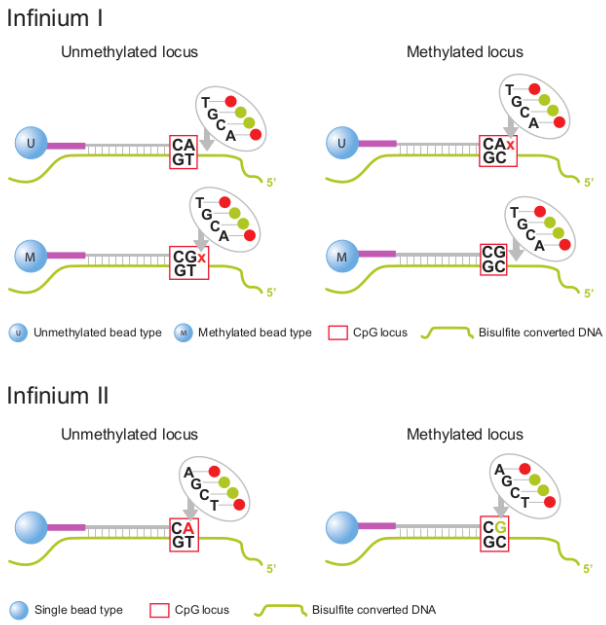

Figura 4: Cobertura mais ampla usando projetos de ensaio Infinium I e II - O MethylationEPIC BeadChip utiliza os ensaios Infinium I e Infinium II. O Infinium I emprega 2 sondas por locus de CpG (cada CpG tem 2; metilado e não metilado), 1 para U e outro para M. Dessa forma, cada CpG tem 4 sondas. O design do Infinium II utiliza uma sonda por locus de CpG, isto é, 2 sondas por ilha.(Illumina 2015)

5 Análise de Dados de Metilação

5.1 Pré-processamento

Os dados coletados no chip. . . .

O pacote Bioconductor oferece diversas ferramentas de pré-processamento

5.1.1 Métodos Iterativos

5.1.2 Métodos Estatísticos

5.2 Análise de qualidade

6 Métodos de Aprendizado de Máquina não Supervisionado

6.1 Redução de Dimensões

6.1.1 PCA

6.1.2 TSNE

6.1.3 UMAP

6.2 Clustering baseado em distância

6.2.1 Hierárquico

6.2.1.1 Divisivo

6.2.1.2 Aglomerativo

6.3 Clustering baseado em densidade

6.3.1 DBSCAN

6.3.2 OPTICS

6.4 Clustering baseado em Mistura de Modelos Gaussianos (GMM)

Referências bibliográficas

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.

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

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.

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

Illumina. 2017. "An introduction to Next-Generation Sequencing Technology", 16.

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

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.

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

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.

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