Métodos de aprendizado de máquina supervisionado para classificações aplicados a dados de RNA-Seq

Resumo

O sequenciamento de RNA, conhecido como RNA-Seq, é uma tecnologia de sequenciamento *next generation* que utiliza-se do ácido ribonucleico (RNA) para sequenciamento, no qual resulta em informações de expressão gênica de milhares de transcritos simultaneamente. Por meio dela, conjuntamente com métodos de aprendizado de máquina, podemos classificar amostras ou indivíduos em grupos de fenótipos de interesse.

O processo para o sequenciamento inclui a extração do material genético, a seleção do tipo de RNA desejado, a síntese do ácido desoxirribonucleico complementar (DNAc) usando-se do RNA selecionado e, se necessário, subsequente amplificação por meio da cadeia da reação em cadeia da polimerase (PCR). A amostra é levada a um sequenciador de uma préviamente escolhida plataforma. Os dados gerados são gravados de forma bruta e, usualmente, no formato FASTQ, que contém um identificador para a leitura, a sequência lida e o índice de qualidade. Estes dados são posteriormente processados em formato de uma matriz em que, geralmente, as linhas são os genes, as colunas são as amostras e cada elemento é a contagem de genes.

Há diversos métodos documentados de aprendizado de máquina, um exemplo clássico é a Análise Discriminante Linear (LDA), utilizada para a classificação de dois ou mais grupos tendo como suposição dados com distribuição de normal multivariada. No caso de dados originados de RNA-Seq, de natureza discreta, LDA não é adequado. Foram utilizados, portanto, os métodos da Análise Discriminante de Poisson (PLDA) e da Análise Discriminante da Binomial Negativa (NBLDA), mais apropriados para dados discretos.

Depois o estudo teórico desses métodos, os mesmos foram aplicados em um conjunto de dados reais de RNA-Seq, disponíveis publicamente na plataforma *Gene Expression Omnibus* (GEO) da *National Center for Biotechnology Information* (NCBI) no código GSE146889. São amostras de 176 tecidos: 91 tumores e 85 normais; dentre os tumores, 19 estão associados a alta instabilidade de microssatélites devido hipermetilação (MSI-H MLH1), 19 não estão associados a ausência de instabilidade de microssatélites (MSS) e 53 são associados a instabilidade de microssatélites em pessoas com casos putativos de Síndrome de Lynch (MSI-LS).

O software *R*, mais especificamente o pacote *MLSeq*, disponibilizado no *Bioconductor*, foi utilizado no ajuste dos métodos no conjunto de dados. As análises foram feitas de duas formas: a primeira apenas considerando tumores e normais e a segunda considerando as

quatro classes (normais, MSS, MSI-H MLH1 e MSI-LS). Usando o método de validação cruzada, as amostras foram divididas aleatoriamente em 70% para treino e 30% para teste. O código completo utilizado na obtenção dos resultados estão disponíveis em github.com/FrancisAkio/Inicacao-Cientifica.

Quando os métodos foram aplicados considerando apenas duas classes, os três modelos tiveram resultados satisfatórios, sendo o NBLDA o modelo que teve melhor resultado, com uma acurácia de 88,46%. Quando os métodos foram aplicados considerando as quatro classes de tecidos, os modelos NBLDA e PLDA tiveram a mesma acurácia de 65,38%, sendo ligeiramentes melhores do que o método PLDA com transformação de potência, com 63,46% de acurácia; de modo geral, a classificação não foi muito boa, apesar dos tecidos normais e do tipo MSS terem sido razoavelmente bem classificados, os tecidos MSI e MSI-LS, não foram.