XXVIII Congresso {virtual} de Iniciação Científica da Unicamp 2020

Métodos de agrupamento aplicados a dados de microarranjo de DNA

Orientadora: Samara F. Kiihl
Aluno: Caio Henrique de Sousa Lima

Setembro de 2020

Resumo

A tecnologia de microarranjo de DNA torna possível monitorar simultaneamente a expressão de milhares de genes para várias amostras. Extrair padrões genéticos a partir desses dados é um dos objetivos em genômica. No entanto, devido ao grande número de genes, interpretar resultados obtidos a partir desses experimentos é um desafio. Uma abordagem bastante utilizada tem sido empregar metodologias estatísticas de agrupamento de dados, para explorar e identificar estruturas e padrões interessantes nos dados. Neste trabalho, iremos estudar os algoritmos de agrupamento mais comuns e sua aplicação em conjunto de dados reais de microarranjo.

Palavras Chaves: Aprendizado Não Supervisionado, Agrupamento, K-means, Agrupamento Hierárquico, Microarranjo de DNA.

1 Introdução

Os avanços nas pesquisas trouxeram tanto conhecimento quanto novas tecnologias, que na área da biotecnologia resultaram em um enorme progresso. Pesquisas que antes só conseguiam analisar pequenas quantidades de genes por vez, agora geram um grande volume de dados pelo sequenciamento de genomas, que aliado com a quantidade de dados sobre expressão genética, tornou complexo a compreensão das finalidades dos genes nos organismos. Dessa forma, surge

a tecnologia de Microarranjos, técnica que possibilitou a análise de grandes expressões genéticas, através de um experimento simples, rápido e eficaz.

A tecnologia de microarranjo consiste em chips de DNA, contendo amostras de RNA's que quando combinadas com reagentes químicos emitem uma luz fluorescente de acordo com a condição de interesse. Com as imagens geradas, é possível mensurar os dados biológicos (genes) cuja análise tem levado a várias descobertas. Enquanto o aprendizado de máquina é "a capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência", sendo extremamente útil para a análise dos dados gerados pela tecnologia de microarranjo.

Deste modo, o presente trabalho possui o intuito de utilizar de técnicas de aprendizado de máquina não supervisionado, identificando possíveis grupos de observações de acordo com os genes dados. Para isto, será testado se é possível separar pacientes em subgrupos de forma que a diferença entre pacientes com choque séptico e sem choque séptico esteja bem delimitada.

2 Aplicação em Dados Reais

2.1 Sepse e Choque Séptico

A Sepse é uma condição em que uma infecção chega a corrente sanguínea e causa inflamações em outras partes do corpo. É considerada como uma resposta desregulada do sistema inflamatório e imunológico a uma invasão microbiana, tendo uma taxa de mortalidade de 15% a 25%, chegando a produzir lesões a órgãos e produzindo febre (ou hipotermia), taquicardia (aceleramento dos batimentos cardíacos), taquipneia (aumento da frequência respiratória) e mudanças de leucócitos no sangue. O Choque Séptico é a evolução do quadro de Sepse, com sintomas de hiperlactatemia e hipotensão, surgindo a necessidade de introduzir agentes anti-hipotensivos no paciente, e desta forma, a mortalidade passa a ser de 30% a 50%. Existem outros tipos de choques que se manifestam de maneiras diferentes, mas que levam ao mesmo estágio final de falência múltipla de órgãos como resultado do desequilíbrio entre a demanda e fornecimento de oxigênio.

Pacientes submetidos a procedimentos cirúrgicos ficam mais expostos a infecções, que podem atingir um quadro de Sepse ou de Choque Séptico. Porém, atualmente não há um padrão a ser seguido para diagnosticar a Sepse, e, desta forma, há um desafio em diferenciar um choque séptico e um choque não séptico após uma cirurgia, já que os pacientes de ambas as condições apresentam sintomas similares. Neste sentido, é necessário um diagnóstico rápido e preciso de choque séptico, para permitir um tratamento imediato desta condição.

Tendo em mente os desafios estabelecidos, o objetivo deste trabalho é avaliar expressões genéticas de pacientes pós-cirúrgicos com choque séptico e com choque não séptico, realizando técnicas de agrupamento com o fim de obter subgrupos que indiquem uma boa divisão.

2.2 Materiais e Métodos

O conjunto de dados utilizado foi fornecido pelo, em domínio público na plataforma *Gene Expression Omnibus* (GSE) do National Center for Biotechnology Information (NCBI), como número de série GSE131761. Os dados consistem em uma amostra de 129 pacientes, dos quais 81 foram diagnosticado com Choque Séptico pós-cirúrgico, 33 pacientes com Choque Não Séptico pós-cirúrgico e 15 pacientes controle. Ao todo, a expressão genética coletada corresponde a 34127 genes. Os dados utilizados já estavam pré-processados.

Para a execução dos métodos de agrupamento foi utilizado o software *RStudio*. Os dados foram obtidos através das bibliotecas *GEOquery* e *Biobase*, disponibilizadas pelo *Biocunductor*. Os algoritmos dos métodos de agrupamento foram feitos com a biblioteca *stats*, e para a visualização dos resultados foi usado o pacote *kableExtra* para tabelas, e o pacote *factoExtra* para gráficos. Por fim, foi necessário o uso do pacote *genefilter* para a filtragem dos genes.

2.3 Resultados

2.3.1 Filtrando Genes

É comum na literatura realizar um filtro de genes que seriam mais relevantes, que poderiam contribuir para a divisão desejada. Primeiramente, foi calculado o Coeficiente de Variação para cada gene. Como não existe uma regra que estabeleça a partir de qual medida os genes são filtrados, foram testadas diversas possibilidades cortes a partir do quantil do conjunto de dados. Conjuntamente, para cada filtro testado, foram realizados testes de hipóteses para os genes restantes e que tiveram seus respectivos p-valores ajustado por FDR. Além disso, para cada filtro, foi testado diferentes níveis de significância para a escolha dos genes pelos testes de hipóteses.

Considerando uma linha de corte de 24% para o quantil e a um nível de significância de 0.01, chegou-se em 154 genes finais, pois assim obteve-se o melhor resultado, usando Cluster Hierárquico, que pode ser visualizado na Figura 1. Ao invés de obter 3 clusters, verificou-se que ao separar um dos subgrupos o resultado ficou mais interessante. A Tabela 1 mostra que os subgrupos 1 e 2 são compostos majoritariamente por pacientes com Choque Séptico, e, analogamente, o subgrupo 3 é composto por paciente com Choque Não Séptico em sua maioria, restando o último subgrupo com apenas pacientes controle.

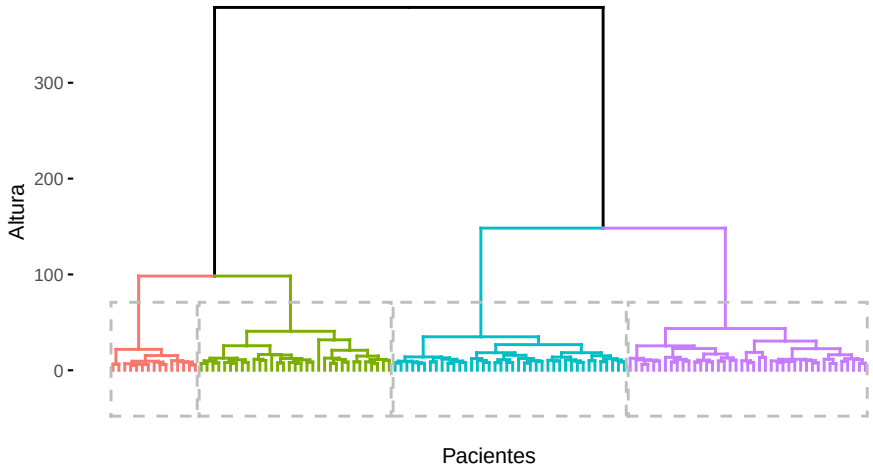

Figura 1: Dendograma com 4 Clusters após a filtragem de genes

Tabela 1: Resultados por Cluster Hierárquico com Genes Filtrados

	Cluster			
Diagnóstico	1	2	3	4
Choque Não Séptico	0	7	26	0
Choque Séptico	41	33	7	0
Controle	0	0	0	15

Os resultados obtidos acima são completamente reprodutíveis e a análise completa pode ser encontrada na plataforma *GitHub*, pelo link: https://github.com/CaioHSLima/Proj-IniciacaoCientifica-CNPq.

3 Discussão e Considerações Finais

O objetivo do presente trabalho foi compreender e aplicar métodos de aprendizado não supervisionado, voltado para o agrupamento de dados de microarranjo de DNA, com fim de encontrar subgrupos de interesse. Para isto, foram escolhidos dados com expressões genéticas de pessoas que tiveram Choque Séptico ou outro tipo de Choque, e assim, usou-se o k-means e o cluster hierárquico como métodos de agrupamentos na tentativa de separar os dois grupos pelas expressões genéticas.

Foi abordado o que é o microarranjo de DNA e sua importância, bem como alguns conceitos biológicos. Em conjunto com os métodos de aprendizado de máquina não supervisionado, pode-se agrupar pessoas em diferentes condições de acordo com suas expressões genética, o que poderia auxiliar em um diagnóstico mais preciso e rápido.

Ao aplicar ambos os métodos nas expressões genéticas, ocorreu um confundimento entre os grupos de pessoas com Choque Séptico e Choque Não Séptico, ou seja, nos clusters formados os grupos ficaram misturados, com exceção dos pacientes controle. Em seguida, reduziu-se o número de variáveis através do PCA, que indicou componentes principais que mais explicam a variabilidade dos dados. Porém, agrupar utilizando as componentes principais obteve resultados similares que anteriormente, com confundimento entre grupos.

A presença de tantos genes podem ter contribuído para a má divisão, visto que há muitos com baixa variância e acabam não se diferenciando entre os grupos de interesse. Tendo isso em mente, foi proposta uma filtragem dos genes, com o intuito de utilizar aqueles que mais se diferenciavam e contribuíssem em agrupamentos mais adequados. Primeiramente filtrou-se os genes que possuíam um baixo coeficiente de variação, que estavam abaixo da linha de corte de 24% do quantil do conjunto de dados. Em seguida, realizou-se teste de hipóteses gene por gene, para identificar os que são diferentes entre as categorias. Por ser uma grande quantidade de testes, ajustou-se os p-valores obtidos com FDR para reduzir a taxa de falsos positivos. A um nível de significância de 0.01, chegou-se em 154 genes finais. O método de cluster hierárquico se mostrou melhor, que agrupou os dados em 4 grupos: dois compostos majoritariamente por pessoas com Choque Séptico, um por Choque não Séptico em sua maioria, e o último apenas por pessoas controle.