Modelos Estatísticos Aplicados à Biologia Computacional e Medicina de Precisão

Palavras-chave: BIOESTATÍSTICA, GENÔMICA, APRENDIZADO ESTATÍSTICO DE MÁQUINA

Autores/as:
Bruna Caroline Dias Silva[Universidade Estadual de Campinas]
Prof. Dr. Benilton de Sá Carvalho (Orientador) [Universidade Estadual de Campinas]
Co-autores/as:
Mariângela Lima Rodrigues[Universidade Estadual de Campinas]

Sobre o projeto

A Estatística é uma área fundamental que envolve todo o processo de análise de dados de modo a obter informações cruciais a partir destes a fim de que se possa tomar decisões de forma assertiva. No que tange a área da medicina é possível observar as mais diversas aplicações que podem vir a afetar de forma direta a vida de um paciente, desde o compreendimento de sua genética até o entendimento e tomada de decisões quanto ao tratamento a ser aplicado de forma a intervir em algo que lhe causa malefícios, por exemplo.

Neste sentido, o projeto desenvolvido visa tratar dos problemas envolvidos no processo de análise de dados de alto-rendimento obtidos através de microarranjos, no que tange a área da Biologia Computacional, Estatística e Medicina de Precisão, tendo como objetivo principal a aplicação de modelos estatísticos para a quantificação da expressão gênica.

1. Introdução à Biologia Molecular e Microarranjos

O conceito de hereditariedade está relacionado aos processos biológicos que permitem com que a informação genética seja transmitida para o sucessor por meio da reprodução. Deste modo, as características biológicas dos seres vivos são moldadas com base na informação genética armazenada no núcleo das células dentro do que é conhecido como DNA (ácido desoxirribonucleico).

O DNA é um composto orgânico formado por fitas de dupla hélice, estas compostas por açúcar, fosfato e bases nitrogenadas que, em diferentes combinações, geram todas as características biológicas do ser vivo. Este também é relacionado com enfermidades que podem surgir ao longo da vida, sendo um importante objeto de estudo na identificação de genes (sequências de DNA) relacionado a alguma característica de interesse, como por exemplo uma doença.

A medicina de precisão é a área que tem por objetivo individualizar o tratamento médico do paciente com base em informações genéticas e características biológicas. Para isso são utilizadas diversas técnicas diferentes, sendo os Microarranjos uma das ferramentas mais empregadas no processo.

Microarranjos são pequenas placas feitas de vidro ou acrílico que são divididas de maneira matricial e nelas estão contidas moléculas de DNA sintéticas, denominadas sondas. Um experimento que utiliza placas de microarranjos é composto por diversas etapas, como a fragmentação e amplificação de amostras de DNA, rotulagem e hibridização da amostra, e a quantificação da molécula-alvo em cada sonda que é feita através da análise de fluorescência.

Essa técnica de análise genética permite observar e estudar o polimorfismo de nucleotídeo único (SNP) - um tipo de variação genética mais comum entre pessoas - possibilitando a identificação de variantes de DNA, a detecção de genes cujas expressões caracterizam doenças e etc.

Salienta-se que cada etapa do experimento com microarranjos pode influenciar na precisão da transcrição do gene, assim é indispensável tanto para o experimentalista, quanto para o analista dos dados, conhecer cada etapa precisamente. Um detalhamento dessas etapas é apresentado por Jaksik, R., Iwanaszko, M., Rzeszowska-Wolny, J., & Kimmel, M. (2015).

Tendo em vista o contexto apresentado, este projeto tem por objetivo estudar as técnicas estatísticas envolvidas na modelagem da quantificação da expressão gênica do indivíduo a partir de dados de microarranjos.

2. Pré Processamento

O processo de análise de dados de microarranjos consiste em quantificar a intensidade da luz que é refletida pelas células do microarranjo devido ao processo de hibridização e, após a quantificação analisam-se os dados obtidos de tal modo que se possa dizer sobre questões de interesse.

Entretanto, é necessário observar que durante a quantificação dos dados dos microarranjos, assim como qualquer outro processo de coleta de dados, existem interferências externas que podem vir a comprometer a precisão dos dados. Tendo isso em vista, antes que os dados sejam processados e a análise seja feita, considera-se extremamente necessária uma etapa anterior às análises que é denominada de pré processamento.

O pré processamento compreende todo o processo de tratamento dos dados quantificados, que pode ser resumido em três etapas principais (Bolstad, B. M. (2004)) - Correção de ruído de fundo; Normalização dos dados e Sumarização - de modo a reduzir os efeitos dos fatores externos sobre os dados obtidos tornando-os mais precisos e com melhor qualidade. Todas estas etapas são realizadas através do pacote `oligo` disponibilizado pela comunidade Bioconductor.

3. Análise de Dados

Foi analisado um conjunto de dados com respeito a 200 pacientes dos quais 183 são pacientes que possuem síndrome mielodisplásica (MDS) e 17 não possuem, estes últimos são denominados pacientes controle (CTRL). A fim de simplificação, o banco de dados foi reduzido para 34 observações, das quais 17 são do tipo MDS e 17 observações são do tipo CTRL. As amostras foram obtidas a partir da medula óssea de cada paciente e os dados foram disponibilizados por Pellegatti et al. (2010) no repositório GEO, sob o identificador GSE19429.

Os dados foram submetidos ao processo de pré processamento conforme descrito na seção anterior através do uso da função `rma()`, que utiliza a metodologia Robust Multiarray Average (RMA), disponível no pacote `oligo` da comunidade Bioconductor. Este processo foi feito utilizando o software `R Studio 4.0.3`.

A partir disso foi possível sumarizar os dados, e estes foram modelados usando Linear Models for Microarray Data, ver mais em (Ritchie et al., 2015). Este modelo pode ser visto na Equação 1, onde a variável resposta yij indica a intensidade esperada em escala logarítmica para o gene i do indivíduo j e o índice k representa a variável preditora, que, para esse estudo em particular, é uma variável indicadora associada ao indivíduo pertencente ao grupo MDS.

$$y_{ij} = \beta_{i0} + \sum_{k=1}^{p} \beta_{ik} x_{jk} + \epsilon_{ij} \tag{1}$$

O ajuste do modelo foi realizado com o auxílio do pacote `limma` (Ritchie et al., 2015).

Feito os ajustes para cada alvo do microarranjo, é de interesse analisar a intensidade das estimativas dos coeficientes $\hat{\beta}$ para assim determinar quais os genes são significativamente diferentes entre os grupos, assim

relaciona-se $\hat{\beta}_0$ ao efeito da log-intensidade do grupo de indivíduos denominados controle, ao passo que a log intensidade para o grupo acometido pela síndrome mielodisplásica é dada acrescentando o coeficiente $\hat{\beta}_1$. Para encontrar os genes candidatos que possuem expressões significativamente diferentes pode-se olhar para a diferença expressas entre os grupos (logFC) e a taxa de falsos positivos, definindo como possíveis candidatos aqueles que estão abaixo do ponto de corte.

A partir disso, foi possível encontrar os 10 genes que mais evidenciam diferença no nível de expressão, estes são apresentados na Tabela 1 abaixo.

Tabela 1: Os 10 candidatos com maiores evidências de expressão diferencial.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
208285_at	-2.03	5.62	-10.75	0	0	17.95
1007_s_at	-0.95	6.74	-8.95	0	0	13.71
1559716_at	-1.07	6.40	-8.87	0	0	13.52
219976_at	1.47	6.17	8.38	0	0	12.28
205640_at	0.98	6.74	7.97	0	0	11.20
219615_s_at	1.12	6.56	7.88	0	0	10.97
1556599_s_at	-2.58	5.35	-7.70	0	0	10.49
1555779_a_at	-1.02	6.16	-7.59	0	0	10.20
230015_at	-0.89	4.94	-7.51	0	0	9.99
220157_x_at	1.02	7.23	7.49	0	0	9.92

Para um melhor mapeamento dos genes que expressam uma diferença significativa entre os grupos, foi utilizado o gráfico vulcão (Neves, 2010) apresentado pela Figura 1, onde o eixo x representa a diferença expressa pelos dados do microarranjo (logFC) e o eixo y os valores descritivos (P-valores). Deste modo, os genes que mais expressam diferenças entre sí são os pontos que possuem, em módulo,um valor alto tanto para o logFC como para o negativo do log do P-valor. O uso de tal técnica é interessante pois, além de facilitar a identificação dos genes que apresentam diferença entre os grupos, ainda permite visualizar resultados que levariam a falsas descobertas, sendo estes os pontos cujo módulo de logFC é alto e o negativo do log do P-valor é baixo.

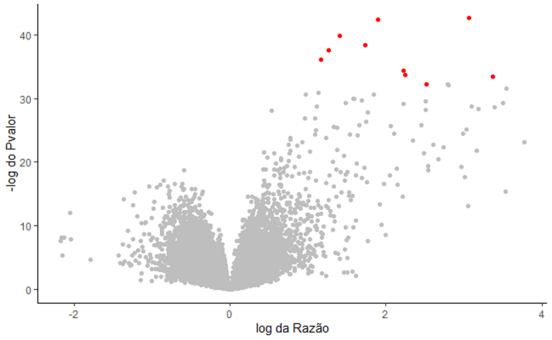

Figura 1: Gráfico vulcão

Além disso, foi possível estudar, para os 10 genes selecionados, a intensidade expressa após o pré-processamento para cada indivíduo. A ferramenta utilizada foi o mapa de calor (ver mais em Khomtchouk, B. B., Van Booven, D. J., & Wahlestedt, C. (2014)) apresentado na Figura 2, onde as linhas apresentam os indivíduos, e as colunas os 10 genes selecionados. Ainda foi adicionado o status do indivíduo (MDS, CTRL), que permite identificar padrões da expressão em indivíduos que possuem a síndrome mielodisplásica e os que não possuem.

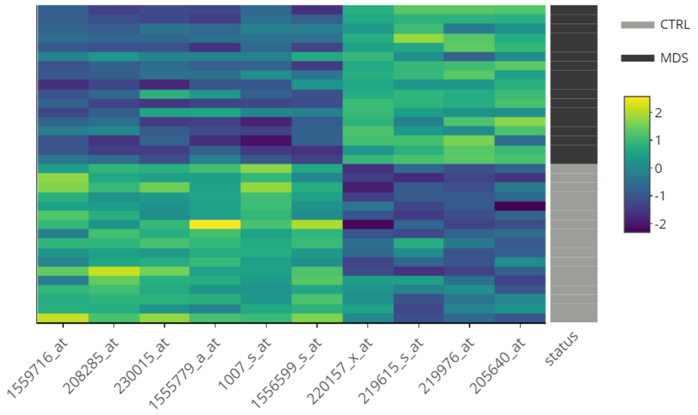

Figura 2: Mapa de calor indicando o potencial de discriminação entre controles e pacientes de MDS. Cada coluna representa um alvo no microarranjo e cada linha indica um indivíduo diferente.

Deste modo foi possível aplicar as técnicas estudadas para a análise de dados no campo da biologia computacional e medicina de precisão, verificando tanto a importância do pré-processamento de dados para a obtenção de resultados confiáveis, uma vez que este foi capaz de reduzir interferências externas, quanto o bom desempenho o modelo linear empregado, sendo este capaz de identificar os genes que expressam diferenças significativas entre o grupo MDS e controle.

6. Bibliografia

Bolstad, Benjamin Milo. 2004. Low-Level Analysis of High-Density Oligonucleotide Array Data: Background, Normalization and Summarization. University of California, Berkeley Berkeley.

Jaksik, Roman, Marta Iwanaszko, Joanna Rzeszowska-Wolny, and Marek Kimmel. 2015. "Microarray Experiments and Factors Which Affect Their Reliability." Biology Direct 10 (1): 1–14.

Khomtchouk, Bohdan B, Derek J Van Booven, and Claes Wahlestedt. 2014. "HeatmapGenerator: High Performance Rnaseq and Microarray Visualization Software Suite to Examine Differential Gene Expression Levels Using an R and C++ Hybrid Computational Pipeline." Source Code for Biology and Medicine 9 (1): 1–6.

Neves, Carlos Eduardo. 2010. "Experimentos de Microarrays E Teoria Da Resposta Ao Item." PhD thesis, Universidade de São Paulo.

Pellagatti, A, M Cazzola, A Giagounidis, J Perry, L Malcovati, MG Della Porta, M Jädersten, et al. 2010. "Deregulated Gene Expression Pathways in Myelodysplastic Syndrome Hematopoietic Stem Cells." Leukemia 24 (4): 756–64.

Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. 2015. "limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." Nucleic Acids Research 43 (7): e47. https://doi.org/10.1093/nar/gkv007.

Sousa, Heidi Mara do Rosário, and others. 2019. "Estudo de Modelos de Classificação Com Aplicação a Dados Genômicos."