APLICAÇÃO DO CRITÉRIO DE FDR (False Discovery Rate) PARA O CONTROLE DE FALSOS POSITIVOS EM UM EXPERIMENTO COM MICROARRAYS

Natália Faraj Murad[1], Rosiane Rodrigues Alvez[1], Júlio Sílvio de Sousa Bueno Filho[1].

1- Universidade Federal de Lavras, Departamento de Ciências Exatas.

Resumo: Os microarrays são uma importante ferramenta para o estudo de quando e onde os genes são expressos, pois permitem o monitoramento simultâneo da expressão de milhares de genes. A análise estatística de experimento de microarrays é complexa devido à magnitude do experimento que faz com que o erro tipo I aumente com o número de hipóteses testadas. O critério de FDR (False Discovery Rate) proposto por Benjamini & Hochberg (1995) expressa a proporção esperada de hipóteses nulas rejeitadas erroneamente. Essa estimativa possibilita o controle dos falsos positivos. Foi obtida a análise de variância para cada sonda, normalização dos dados e aplicou-se o controle dos falsos positivos. O método de controle da FDR reduziu a proporção de falsos positivos, limitando o número das variáveis-resposta em estudo. A normalização não foi eficiente em todas as sondas.

INTRODUÇÃO

Uma das ferramentas mais usadas atualmente para o estudo de quando e onde os genes são expressos, são os microarrays, que permitem o monitoramento simultâneo da expressão de milhares de genes. A análise estatística para cada sonda é relativamente simples, mas se torna complexa quando é considerado um conjunto de dados de microarrays (milhares de sondas) devido aos testes múltiplos (Dudoit, 2002). Além da magnitude do experimento que envolve alguns milhares de genes e intensifica o grau de erros, geralmente há poucos arrays disponíveis devido ao seu alto custo e, como resultado tem-se um experimento com muitas variáveis-resposta e poucas unidades experimentais (Pereira, 2008).

O primeiro passo da análise estatística consiste na normalização dos dados. Essa transformação é feita após os dados já terem sofrido a transformação padrão logaritmo da intensidade luminosa e tem como objetivo ajustar os efeitos decorrentes da variação na matriz da tecnologia de microarray ao invés das diferenças biológicas entre as amostras de RNA, ou entre as sondas impressas (Smyth & Speed, 2003).

Dentre as razões para que a transformação seja feita estão: colocação de quantidades diferentes de mRNA inicial, diferenças de eficiência de detecção do marcador utilizado, e erros sistemáticos ao medir os erros de expressão (Broche, 2003). Dessa forma, os dados não obedecem às pressuposições de normalidade (distribuição normal, variância constante e independência de sua média) e o modelo análise de variância comum não se ajusta de maneira ideal.

Box & Cox (1964) propuseram um tipo de transformação para dados que não obedeciam as pressuposições de normalidade e essa transformação tem sido amplamente utilizada para normalização de dados de diversos tipos de experimentos. Pode ser aplicada a regressões, numa combinação delas ou a variáveis dependentes numa regressão fazendo com que os resíduos da regressão sejam mais homocedasticos ou mais próximos de uma distribuição normal. A transformação é baseada em uma função de verossimilhança que faz o ajuste dos dados através da expressão:

$$Y^{\lambda} = \begin{cases} \log(Y) \text{ se } \lambda = 0 \\ \dfrac{Y^{\lambda} - 1}{\lambda} \text{ se } \lambda \neq 0 \end{cases}$$

Onde **λ** é o valor que se procura estimar, pois a partir dele os dados transformados são encontrados e então é possível fazer a análise de variância. **Y** é a variável resposta a ser transformada.

Um dos objetivos de ensaios com microarrays é a identificação de sondas com expressão gênica diferencial que são aquelas cujo nível de expressão está associado a uma resposta ou variável de interesse (Dudoit, 2002). Geralmente, nessa etapa é testada uma hipótese para cada sonda e então tem-se um grande número de hipóteses sendo testadas simultaneamente e como conseqüência ocorre o problema de multiplicidade.

Quando muitas hipóteses são testadas, a probabilidade de que o erro tipo I seja cometido aumenta significativamente com o número de hipóteses. O erro tipo I, também chamado de falsos positivos, é o erro que se comete ao rejeitar a hipótese nula quando ela é verdadeira (Benjamini & Hochber, 1995). Comete-se o erro do tipo I ao declarar que uma sonda apresentava expressão gênica diferencial quando na verdade, isso não ocorria (Reiner, 2003).

Benjamini & Hochberg (1995) propuseram controlar a FDR (*False Discovery Rate*), definida como a proporção de hipóteses nulas H_0 verdadeiras, entre as hipóteses nulas rejeitadas, ou seja, a proporção de erros devido à rejeição errônea de H_0. A técnica da FDR vem sendo empregada com sucesso em microarrays, devido ao grande número de variáveis respostas, o que leva à necessidade de reduzir o número de conclusão de falsos positivos para poder manejar, em experimentos posteriores, apenas as variáveis mais promissoras (Pereira, 2008).

Considerando-se uma família de m hipóteses nulas simultaneamente testadas em que m_0 são verdadeiras, para cada hipótese H_i um teste estatístico é calculado com um valor p (P_i) correspondente. Considere que R denota o número de hipóteses rejeitadas por um procedimento, V o número de hipóteses nulas verdadeiras rejeitadas, e S o numero de hipóteses falsas rejeitadas. Considere também que Q é o mesmo que V/R quando R é maior e diferente de zero, pois quando Q = 0 não é cometido o erro de rejeição de hipóteses nulas verdadeiras. Então a FDR é definida como:

$$FDR = E(Q)$$

MATERIAL E MÉTODOS

Os dados utilizados foram disponibilizados no 15º Genetic Analysis Workshop que ocorreu em novembro de 2006. Os microarrays foram obtidos a partir de células linfoblastóides do tipo B de 14 famílias, utilizando-se Affymetrix Human Focus Array que contem sondas para 8500 transcritos. Para 3554 das 8500 sondas testadas, Morley et al (2004) encontraram maior variação dentre os indivíduos que para segmentos replicados num mesmo indivíduo. Essas 3554 expressões fenotípicas foram escolhidas pelo GAW para análises.

Das 14 famílias, duas possuem 13 e doze, 14 indivíduos. O experimento possui 3554 sondas que indicam os níveis de expressão gênica de cada indivíduo, num total de 194. O arquivo de dados encontra-se organizado em colunas, sendo a primeira referente à família, a segunda ao sexo do indivíduo, representado por 1 (masculino) ou 2 (feminino), a terceira ao número do indivíduo na família e as restantes às sondas.

O conjunto de dados sofreu a transformação de Box-Cox para que se ajustassem às pressuposições de normalidade. A transformação foi feita através da função boxcox( ) do pacote MASS no software R v2.10.0 (R Development Core Team, 2010). Essa função estima por máxima verossimilhança os parâmetros de uma transformação de Box-Cox para cada variável. Essa estimativa é feita a partir do gráfico de normalidade de Box-Cox que é um gráfico dos coeficientes de correlação entre os eixos vertical e horizontal do gráfico de probabilidade para vários valores do parâmetro λ. O valor de λ correspondente a máxima correlação no gráfico é escolhido.

A análise de variância foi feita para cada sonda individualmente buscando a identificação de genes diferencialmente expressos, através de um teste de hipóteses. Visou-se avaliar a variável resposta em função da família a qual o indivíduo pertencia e sexo do indivíduo. A partir dessa análise foram obtidos valores de probabilidade pelo teste F, inicialmente a 5% de significância. O procedimento foi aplicado novamente com o teste F a 1% de significância.

O modelo estatístico aplicado para cada sonda foi:

$$Y_{ij} = \mu + \tau_i + \varepsilon$$

em que, Y_{ij} representa as intensidades de expressão em escala logarítmica.

μ é o efeito constante (média geral);

$\boldsymbol{\tau_i}$ é o efeito do i-ésimo tratamento (efeito de sexo e família);

$\boldsymbol{\varepsilon_{ij}}$ é o erro associado ao i-ésimo tratamento na j-ésima unidade experimental ou parcela.

Para aplicação do controle da FDR, os valores do teste F *(P)* das 3554 sondas *(m)* foram ordenados em ordem crescente e então foi obtida sua posição *(i)*. Assim calculou-se um p-valor corrigido através da fórmula:

$$q^* = \frac{P_{(i)}.m}{i}$$

Os algoritmo para realização dos cálculos foi feito através do software R v2.10.0 (R Development Core Team, 2010).

RESULTADOS E DISCUSSÃO

Com a aplicação do controle da FDR espera-se que o número de hipóteses nulas rejeitadas diminua devido à exclusão dos falsos positivos, ou seja, eliminação do erro tipo I. Assim é possível selecionar somente as sondas com maior probabilidade de apresentarem expressão gênica diferencial com base no valor da probabilidade do teste estatístico de hipóteses. Sem o controle da FDR, 2617 hipóteses nulas haviam sido rejeitadas pelo teste F a 5% de significância. Após o controle, no nível de 5% de significância esse número foi reduzido a 2485. Aproximadamente 19% das hipóteses nulas rejeitadas dentre as testadas eram falsos positivos.

Mesmo com a FDR o número de hipóteses rejeitadas continuou alto para estudos posteriores com as sondas. Assim, aplicou-se novamente o teste de hipóteses com nível de significância de 1%. Deste modo, selecionou-se melhor as sondas com maior probabilidade de apresentarem expressão gênica diferencial. Antes da aplicação do FDR, o número de hipóteses nulas verdadeiras rejeitadas foi de 2127. Após a aplicação desse critério, o número foi reduzido a 1909, ou seja, aproximadamente 10% das hipóteses dentre as rejeitadas eram falsos positivos.

Tabela 1: Número de hipóteses rejeitadas dentre as 3554 hipóteses de nulidade testadas para os dados transformados, considerando 1% e 5% de significância, com e sem aplicação do FDR.

	5%	1%
Com FDR	2485	1909
Sem FDR	2617	2127

Através do teste de Shapiro-Wilk a 5% foram eliminadas as sondas nas quais a transformação de Box-Cox não foi eficiente. Assim, dessas 2127 sondas selecionadas pelo critério de FDR, a normalidade foi aceita para 1058 e assim houve mais uma redução no número e maior seleção das variáveis em estudo.

Um exemplo de análise com uma das sondas mais promissoras pode ser encontrado na tabela abaixo (Tabela 2). Neste caso, nota-se que há forte evidência para o efeito de família, indicando que esta sonda pode apresentar controle genético para a expressão gênica.

Tabela 2: Quadro da análise de variância referente à sonda 202203_s_at, dados transformados.

	GL	SQ	QM	F_c	Pr(>F)	FDR
Família	13	174212	13400,9	14,1110	$2{,}181e^{-21}$	$5.454e^{-22}$
Sexo	1	1	0,6	0,0006	0,9798	
Resíduos	179	169993	949,7			
Total	193	344206				

CONCLUSÃO

O método de controle da FDR reduziu a proporção de falsos positivos, limitando o número das variáveis-resposta em estudo, o que facilitou a escolha das sondas

experimentalmente mais promissoras. A transformação de Box Cox não foi eficiente em todas as sondas, mas o teste de Shapiro-Wilk permitiu selecionar aquelas que obedeciam os critérios de normalidade propiciando assim, estimadores válidos.

REFERENCIAS BIBLIOGRÁFICAS

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false Discovery rate: a pratical and Power-full approach to multiple testing. **Journal of the Royal Statistics Society,** v.57, p. 289-300.

BENJAMINI, Y.; KENIGSBER, E.; REINER, A.; YEKUTIELI, D. FDR adjustments of microarrays experiments. **http://www.math.tau.ac.il/~ybenja/Software/fdrame.pdf** Acesso em: 30/04/2009.

BOX, G. E. P.; COX, D. R. An analysis of transformations. **Journal of the Royal Statistical Society.** Series B (Methodological), Vol. 26, No. 2. (1964), p. 211-252.

BROCHE, E.C. Métodos Estatísticos na Análise de Experimentos de Microarray. USP-IME, 2003. **Dissertação de Mestrado.** 108p.

MORLEY, M.; MOLONY, C. M.; WEBER, T. M.; DEVLLN, J.L.; EWENS, K. G.; SPLELMAN, R.S.; CHEUNG, V. G. Genetic analysis of genome-wide variation in human gene expression. **Nature**, London, v.430, n.7001, p. 743-747, Aug. 2004.

PEREIRA, R. N. 2008. Controle do erro tipo I em um experimento de microarrays com eucalipto. 57p. **Tese (Doutourado em Estatística e Experimentação Agropecuária)** – Universidade Federal de Lavras, Minas Gerais, Lavras.

R DEVELOPMENT CORE TEAM. R: A Language and Environment for Statiscal Computing. Vienna, Austria, 2010: R Foundation for Statistical Computing. Disponível em: **http://www.r-project.org**.

REINER, A.; YEKUTIELI, D.; BENJAMINI, Y. Identifying differentially expressed genes using false discovery rate controlling procedures. **Bioinfomatics**, v.19, n3, p. 368-375. 2003.