ARTIGO ORIGINAL

ISSN 1677-5090

Experimentos com microarrays: visão geral e comparação de modelos estatísticos para a identificação de genes diferencialmente expressos

Microarray experiments: overview and comparison of statistical models for identifying differentially expressed genes

Rejane Augusta de Oliveira Figueiredo[1], Júlia Maria Pavan Soler[1]
[1]*Instituto de Matematica e Estatística da USP (IME-USP).*

Resumo

Grandes avanços estão ocorrendo na área de Genética e Genômica. Inicialmente cada gene era analisado, separadamente, com o intuito de se verificar, por exemplo, associações com desenvolvimento de doenças. Recentemente, surge a técnica de microarrays que permite que milhares de genes sejam avaliados, simultaneamente. Esta técnica apresenta diversas vantagens nas aplicações em muitas áreas do conhecimento como, por exemplo, a área médica, porém uma série de dificuldades são encontradas nestes experimentos devido às diversas fontes de variações sistemáticas que podem interferir nas mensurações obtidas, acarretando em resultados falso-positivos. Devido a estas variações e a outros problemas encontrados nestes experimentos, como o problema de múltiplos testes, pois milhares de genes são avaliados num mesmo momento, muitos são os esforços em se encontrar uma melhor abordagem de análise estatística para a identificação de genes, diferencialmente, expressos (DE). Baseado nestes aspectos, no presente estudo serão apresentadas e comparadas possíveis técnicas de análise estatísticas úteis na identificação de genes DE. Como aplicação e motivação deste trabalho, algumas técnicas de análise são aplicadas a um conjunto de dados reais com ratos diabéticos.

Palavras-chave: Microarrays. Genes diferencialmente expressos. Modelos mistos.

Abstract

Great advances are occuring in the Genetics and Genomics area. At the begining each gene was analysed separatedly with the intention to verify, for example, associations with the development of diseases. Recently the microarrays technique seems to allows that thousands of genes are evaluated simultaneously. This technique presents several advantages in applications in many areas, such as the medical area. However, a series of difficulties are found in these experiments due to several sources of systematic variations that can interfere in the results, and may come to false-positive results. Due to these variations and other problems found in these experiments, such as the multiple tests problem, because thousands of genes are evaluated at the same moment, many are the efforts at finding a better approach of statistical analysis for the identification of differentially expressed genes (DE). Based on these aspects, this study presents and compare possible techniques of statistical analysis useful in the identification of the genes DE. As application and motivation of this work some techniques of analysis are applied to a set of real data with diabetic mice.

Key Words: Microarrays. Differentially expressed genes. Mixed models.

INTRODUÇÃO

Conhecer a estrutura genética possibilita trazer informações de características dos indivíduos, animais e vegetais, com forte impacto em áreas de aplicações das ciências biológicas, como a medicina e a agricultura. Nos dias de hoje, diversas técnicas estão sendo utilizadas para entender a estrutura genética de uma grande variedade de doenças com o objetivo de produzir melhores diagnósticos, prevenção e cura destas doenças (Satagopan; Panageas, 2003).

Os mapas genéticos, mapas físicos e sequenciamento tratam do genoma estrutural. A partir das informações obtidas no genoma estrutural, começa o interesse pela genômica funcional que tem como objetivo compreender como e quando os genes são expressos em um organismo, quais suas inter-relações com outros genes e com o ambiente, e como são regulados. Sendo assim, a genômica funcional, usando as informações geradas pela genômica estrutural, leva à completa caracterização do padrão de expressão do conjunto completo dos genes, assim como à investigação sistemática das propriedades funcionais desse conjunto de genes (Simpson; Caballero, 2000).

É no campo da genômica funcional que surgem as técnicas para avaliar e entender a expressão de alguns genes em determinadas situações. *Microarrays* é uma das técnicas utilizada para medir e investigar a expressão genética, que tem como grande vantagem a possibilidade de avaliar uma enorme quantidade de genes ao mesmo tempo (Draghici, 2003). A análise das informações encontradas, nesta e nas demais técnicas

Recebido em 11 de junho de 2011; revisado em 16 de agosto de 2011.
Correspondência / Correspondence: Rejane Augusta de Oliveira Figueiredo. Rua Ouvidor Peleja, 585. apto. 74. Bairro Vila Mariana. CEP: 04128-001. São Paulo - SP. Email: rejanef@uol.com.br

disponíveis, é realizada utilizando-se métodos computacionais e estatísticos intensivos e não triviais. Neste cenário, é crescente a necessidade de desenvolvimento e aperfeiçoamento de ferramentas computacionais específicas, bem como de suas aplicações em novas abordagens experimentais.

Com o avanço biotecnológico, abordagens que avaliavam em um experimento um único gene ficaram ultrapassadas e, assim, foram desenvolvidas novas estratégias capazes de fornecer uma perspectiva global e simultânea da expressão do genoma de um organismo ou tecido por meio de um único experimento. Dentre estas estratégias está a tecnologia de *Microarrays*, que foi desenvolvida para explorar os dados da seqüência de DNA e produzir informações sobre o nível de expressão gênica do genoma (Dudoit et al., 2000).

A técnica de *microarrays* é utilizada para medir a expressão de milhares de genes, através da imobilização (fixação) de fragmentos previamente conhecidos de DNA em uma lâmina. Utilizando-se o princípio da complementaridade, o mRNA dos diferentes tratamentos de interesse irá se juntar à seqüência fixada na lâmina. Há diversos tipos de técnicas, porém as mais utilizadas atualmente são: a técnica de *microarrays* baseada em oligonucleotídeos[1], onde curtos pedaços da seqüência de DNA são pré-fixadas nas lâminas, e a técnica baseada no DNA complementar (cDNA), onde longas seqüências de DNA são fixadas na lâmina. Além desta diferença, há outras diferenças entre tais técnicas, a citar: nos experimentos com oligonucleotídeos para cada tratamento de interesse no estudo é utilizada uma diferente lâmina, já nos experimentos com cDNA, em cada lâmina são avaliados dois diferentes tratamentos, simultaneamente (Woo et al., 2004). É difícil julgar qual das tecnologias é superior, porém algumas vantagens e desvantagens podem ser citadas como: nos experimentos com cDNA pode haver uma seqüência de DNA fixada na lâmina que seja desconhecida, o que, em geral, não ocorre em experimentos com oligonucleotídeos; pode haver uma maior fonte de variação sistemática nos estudos com cDNA como, por exemplo, variação devido às diferentes colorações usadas para definir os tratamentos dentro de cada lâmina, isto não ocorre nos experimentos com oligonucleotídeos, pois em cada lâmina há apenas um tratamento com apenas uma coloração de identificação; uma vez que tenha sido feito um bom planejamento do experimento e que haja interesse na comparação de apenas dois tratamentos a análise com cDNA é mais simples; além destes, o custo das lâminas de oligonucleotídeos são mais caras (Draghici, 2003).

Para as duas técnicas, tanto com oligonucleotídeos quanto com cDNA, uma seqüência de uma única fita do DNA complementar de um gene de interesse é fixada nas cavidades (do inglês *spots)* presentes nas lâminas de materiais como vidro, nylon ou folha de quartzo *(*Huber; Von Heydebreck; Vingron, 2003). As amostras de mRNA para cada tratamento de interesse são marcadas com uma determinada fluorescência. Em experimentos com oligonucleotídeos, por conterem estes apenas uma lâmina para cada tratamento avaliado, haverá uma única cor de fluorescência e, em experimentos com cDNA, duas cores de fluorescência, uma para cada tratamento em uma mesma lâmina.

Em estudos com cDNA, cada gene é representado por uma grande e única seqüência de bases. Além disso, como citado anteriormente, para identificação de genes DE, em cada lâmina, amostras de mRNA de dois tratamentos são utilizadas para a análise. As duas amostras são marcadas com fluorescências de diferentes cores, em geral, verde (cianina 3 - denotado por Cy3) e vermelha (cianina 5 - denotado por Cy5), e depois hibridizadas[2] na lâmina, onde as seqüências de cDNA estão fixadas. O presente estudo será baseado nesta técnica de cDNA, com duas amostras (tratamentos) em cada lâmina.

Embora os dois tipos de lâminas (*arrays*) citados, com oligonucleotídeos e com cDNA, produzam expressão gênica usando diferentes métodos, ambas podem ser dirigidas para uma pesquisa com a mesma finalidade em questão e técnicas estatísticas similares de análise dos dados podem ser aplicadas (Satagopan; Panageas, 2003).

As medidas de expressão gênica obtidas por meio das técnicas de *microarrays* podem apresentar diversas fontes de variação, como problemas na preparação das amostras de mRNA de cada tratamento antes de serem hibridizadas, erros devido às propriedades de hibridização da lâmina, falhas na detecção da fluorescência e até mesmo sujeira na lâmina (Holder et al., 2001). Como exemplo, em estudos com mais de uma lâmina, pode haver variação na qualidade de uma lâmina para outra e, assim, devido a condições inconsistentes do experimento, pode-se aumentar ou reduzir a eficiência de hibridização de um cDNA, fazendo com que as diferenças de intensidade da expressão gênica ocorram devido às diferentes lâminas e não aos diferentes tratamentos. Além disso, uma das colorações pode estar mais brilhante do que a outra por propriedades físicas ou químicas. Desta forma, um gene pode emitir um sinal de fluorescência mais alto (ou mais baixo) do que outros genes por apresentar um maior nível de expressão, ou devido a estas variações na eficiência da hibridização e/ou na eficiência da coloração das amostras de mRNA. Outro grande problema encontrado neste tipo de estudo é que uma enorme quantidade de genes são avaliados, porém com uma baixa quantidade de unidades experimentais

[1] Oligonucleotídeo – fragmento sintético de DNA composto por somente poucas seqüências (em torno de 25 pares de bases nitrogenadas).
[2] Hibridizar – é a ligação de seqüências de DNA e/ou RNA através do pareamento das bases nitrogenadas complementares para formar uma cadeia de dupla fita.

(Draghici, 2003), isso devido, principalmente, ao alto custo das lâminas. Como exemplo, os dados utilizados neste estudo contêm informação de apenas 24 animais e para cada animal foram avaliados 12.768 genes. Vários são os esforços realizados em experimentos com *microarrays*, visando minimizar as variações encontradas nestes experimentos devido a fatores externos e indesejáveis. Para se garantir a qualidade das medidas de expressão, há os procedimentos de controle *a priori,* realizados por meio de um cuidadoso planejamento do experimento; e os procedimentos de controle *a posteriori,* como as técnicas de normalizações de dados, que são utilizadas após medida a intensidade da expressão gênica.

O presente estudo tem como objetivo apresentar alguns métodos de análise estatística de dados de *microarrays* de cDNA para identificar genes DE. Serão considerados e comparados modelos com diferentes abordagens utilizando-se diferentes formas de normalizar os dados, frente a diversas possibilidades de variações encontradas neste tipo de experimento.

Alguns detalhes Sobre Experimentos com *Microarrays* utilizando-se Lâminas com cDNA

A técnica de *microarrays* de cDNA se baseia, inicialmente, em selecionar quantidades de mRNA nas condições experimentais de interesse, prepará-las e, assim, avaliar o nível de expressão genético de dois diferentes tratamentos em uma mesma lâmina. Podemos exemplificar a técnica com a comparação de dois tratamentos, sendo que em um deles as amostras de mRNA são extraídas de tecidos de pacientes com alguma doença e o outro de tecidos de pacientes sem a doença. Cada uma destas amostras de mRNA (com e sem a doença) são preparadas com fluorescentes de cores diferentes, em geral, uma com verde (G) e a outra com vermelho (R). As amostras com os corantes são misturadas e derramadas nas lâminas já preparadas com as seqüências de DNA previamente conhecidas para que, com estas seqüências, ocorra a hibridização. Através da leitura de um *scanner* a laser são verificadas as intensidades das cores em cada *spot*. Como cada *spot* representa um gene, para cada um deles verifica-se qual intensidade de cor é mais evidente, ou seja, para qual tratamento o gene está mais expresso. Assim, a expressão genética em cada tratamento é avaliada a partir da intensidade de fluorescência (cores) medida nos *spots* que representam um gene de interesse.

Após a hibridização, é feita a leitura da imagem das cores apresentadas na lâmina, por meio de um *scanner* específico com alta resolução espacial. Após realizada a leitura da imagem o grande esforço é quantificar a intensidade desta imagem. O processamento da imagem é realizado por meios computacionais. Neste estudo o software utilizado foi o Genepix, porém há diversos softwares disponíveis para a quantificação e análise das imagens (Speed, 2003). Há diversas opções para a quantificação da imagem e há ainda diversas pesquisas nesta área de análise da imagem visando obter melhores resultados e estimação da intensidade da cor (Speed, 2003).

As intensidades de cada uma das cores medidas pelo *scanner* podem variar numa ordem de grandeza de 0 a 2^{16} (65.535). Para obter a medida de intensidade de expressão, cada grupo de *spots* precisa ser identificado e assim a intensidade quantificada. Vale citar que em cada *spot* a leitura de intensidade de imagem é feita em uma resolução de cerca de 2.500 (50x50) pontos (*pixels*) e a medida de intensidade do *spot* é dada por uma medida resumo destes pontos, em geral, a mediana.

A medida da quantidade de cada fluorescente em cada *spot* mostra uma intensidade que está correlacionada à abundância do correspondente RNA transcrito (Huber; Von Heydebreck; Vingron, 2003). Através da avaliação da intensidade de cada uma das fluorescências, pode-se, então, verificar a qual amostra aquele determinado gene, alocado num determinado spot, pode estar associado. Desta forma, o nível de expressão gênico é medido pela intensidade de fluorescência obtida para as cores Verde e Vermelho, podendo-se obter, de forma geral, as seguintes colorações em cada spot:

Spot com coloração verde: Gene diferencialmente expresso para o tratamento colorido com fluorescência verde;

Spot com coloração vermelha: Gene diferencialmente expresso para o tratamento colorido com fluorescência vermelha;

Spot com coloração amarela: Gene sem expressão em nenhum dos tratamentos;

Spot com coloração preta: Gene que não pode ser avaliado (má qualidade da imagem).

Normalizações das Medidas de Intensidade de Expressão Gênica

Há muitas fontes de variação que podem ocorrer em um experimento com *microarrays* que devem ser controladas para se prosseguir com a análise. Entende-se por normalização possíveis formas analíticas de ajustar as variações sistemáticas que ocorrem neste tipo de experimento e que afetam a medida do nível de expressão dos genes avaliados. Desta forma, é necessário que a normalização das intensidades medidas seja feita antes de iniciar a análise do nível de expressão gênica.

Alguns exemplos de variações sistemáticas que podem ocorrer estão descritas a seguir (Speed, 2003; Draghici, 2003):

Problemas que podem ocorrer durante a preparação da amostra biológica como a extração e

[1] *Housekeeping* - genes responsáveis pela manutenção do metabolismo celular que se mantêm ativos em todas as células do organismo.

isolamento da amostra de mRNA, variação na introdução da fluorescência. Estes problemas podem ocorrer devido a, por exemplo, erro no material utilizado (pipeta) e flutuação de temperatura;

Lâminas de má qualidade podem reduzir ou influenciar a eficiência das hibridizações;

Propriedades técnicas de leitura do *scanner;*

Em algumas situações a intensidade de fluorescência vermelha (R) tende a ser mais fraca do que a intensidade verde, ou um dos canais de coloração (*dye*) pode brilhar mais do que o outro, interferindo na relação entre as duas intensidades de cores.

Em um experimento de *microarrays* utilizando-se cDNA, a normalização dos canais de coloração possibilita que seja feito um balanceamento entre as intensidades dos rótulos de coloração (verde e vermelho), permitindo a comparação dos níveis de expressão entre tratamentos. Porém, como visto anteriormente, podem ocorrer variações devido a outras condições do experimento (Yang et al., 2002). Desta forma, as normalizações podem ser feitas entre lâminas, dentro de lâminas e para os *spots* de cada lâmina.

Em geral as normalizações são feitas por meio de métodos não paramétricos como o Ajuste Global e técnicas como a de Lowess e Spline. Para estes métodos não paramétricos, as normalizações são realizadas a partir dos gráficos MA-PLOT (Yang et al., 2002). Este é um gráfico de dispersão das variáveis M x A, podendo ser construído para o total de dados, em que se tem:

$M_{jg} = log_2(G_{jg}/R_{jg})$, é o logaritmo na base 2 da razão das medidas de intensidade do *g*-ésimo gene hibridizado para os tratamentos corados com G e R, respectivamente, na *j*-ésima lâmina. Tal valor representa a variabilidade existente entre ambos os tratamentos;

$A_{jg} = (\log_2 R_{jg} + \log_2 G_{jg})/2$, é a média do logaritmo das intensidades de expressão do gene *g* na lâmina *j* dos dois tratamentos.

Havendo necessidade de normalização dos dados, as transformações são feitas baseadas nos valores de M e, a partir deles, novos valores de M são obtidos, denotados por M*.

Prosseguindo, após se obter os valores M* pela aplicação de algum método de normalização, chega-se em G* e R* que são as medidas de intensidade já padronizadas que serão utilizadas na análise estatística para identificação dos genes DE. Assim, a partir das fórmulas utilizadas para se construir a ordenada e abscissa do MA-Plot, os logaritmos das intensidades normalizadas serão obtidos da seguinte forma (Rosa, 2004):

$\log_2 G^*_{jg} = A_{jg} + M^*_{jg}/2$;
$\log_2 R^*_{jg} = A_{jg} - M^*_{jg}/2$;
onde,

$\log_2 G^*_{jg}$ – $\log_2$ da medida de intensidade G padronizada do *g-ésimo* gene avaliado na *j-ésima* lâmina;

$\log_2 R^*_{jg}$ – $\log_2$ da medida de intensidade R padronizada para o *g-ésimo* gene avaliado na *j-ésima* lâmina;

M^*_{jg} – $\log_2$ da medida de intensidade da razão G*/R* do *g-ésimo* gene avaliado na *j-ésima* lâmina.

Para a utilização dos métodos de normalização que serão apresentados a seguir, em geral, há uma escolha do conjunto de genes que serão utilizados como base para os cálculos. Três abordagens são geralmente utilizadas para escolha do conjunto de genes (Yang et al., 2002):

Todos os genes da lâmina – é usado em situações em que uma pequena quantidade de genes apresenta-se diferencialmente expressa, fato que, frequentemente ocorre em estudos de *microarrays*;

Genes com expressão constante – usa-se apenas um pequeno conjunto de genes que apresentam nível de expressão que não varia significantemente, mesmo sob condições variadas. Estes são os chamados genes *Housekeeping*[3]. É difícil a identificação destes genes que apresentam expressão constante em qualquer experimento, mas é possível localizar genes considerados *Housekeeping temporários*, que não mudam a expressão quando expostos a algumas situações específicas. Uma limitação dos *Housekeeping* é que muitas vezes eles são altamente expressos, e assim não podem ser representativos de outros genes de interesse;

Elementos de controle – nesta abordagem são utilizadas seqüências sintéticas de DNA ou seqüências de DNA de organismos diferentes das amostras de mRNA que estão sendo estudadas, ou seja, das amostras diferentes daquelas alocadas nos *spots*. Estas seqüências são incluídas em igual quantidade, nos dois diferentes tratamentos do estudo e são colocadas nas lâminas para a hibridização.

Após identificarmos o conjunto de genes que serão utilizados na normalização dos dados, a seguir são apresentados métodos de normalização mais comumente adotados para diversas situações de experimentos de *microarrays*.

No estudo a seguir serão utilizados dois métodos de normalização: um modelo utilizando-se o método de normalização não paramétrica pelo ajuste da função de Lowes e dois modelos utilizando-se métodos paramétricos de normalização.

Métodos de Normalização Não Paramétricos: *Normalização dependente da intensidade – ajuste pela Função de Lowess*

O primeiro modelo avaliado no estudo utilizará esta abordagem de normalização. Nesta abordagem a correção em M é feita permitindo a variação entre *spots* e esta variação pode ser observada em um gráfico MA-PLOT. Os valores padronizados de M, isto é M*, obtidos da normalização, são decorrentes de ajustes baseados

em uma função dependente dos valores de A. Estes valores utilizados na normalização serão subtraídos do valor de M, como é mostrado a seguir:

$$M^*_{jg} = M_{jg} - c(A_{jg}) \quad em\ que\ j=1,...,J\ ;\ g=1,...,G.$$

Um dos métodos utilizados para esse tipo de normalização é através da regressão de Lowess, isto é, a função $c(A_{jg})$ é estimada através do ajuste por Lowess (Cleveland; Loader, 1995). Esta função de Lowess trata de um suavizador de dados em um diagrama de dispersão que realiza um ajuste robusto e não é afetado pela pequena quantidade de genes DE que aparecem como valores aberrantes (*outliers)* em um MA-PLOT. Os ajustes são realizados em pequenos subconjuntos de dados, sendo cada um deles formado por frações (f) do conjunto total de dados, sendo 0<f<1 (Yang et al., 2002). Não há um procedimento ótimo para a escolha do parâmetro de suavização *f*, sendo geralmente escolhido de forma empírica.

Métodos de Normalização Paramétricos

O primeiro método de normalização paramétrica foi proposto por Kerr, Martin e Churchil (2000) que adotaram um modelo de análise de variância clássico (ANOVA) que permite uma análise mais sistemática das fontes de variação que podem influenciar as medidas de intensidade de expressão em um experimento de *microarrays*. É modelada a intensidade da medida de expressão levando em consideração possíveis fontes de variação como a lâmina e o canal de cor, bem como o tratamento e os diferentes genes avaliados. Estas variações são tratadas como fatores fixos do modelo. Com a inclusão destas variáveis no modelo, a normalização se faz no próprio modelo em que os resíduos representam os valores "normalizados" da medida de intensidade. Este será o terceiro modelo avaliado no estudo.

Um segundo modelo paramétrico, sugerido por Wolfinger e colaboradores (2001), é um modelo de efeitos mistos que é bastante flexível para modelar possíveis dependências entre as respostas de intensidade de expressão dentro da lâmina. Este será o segundo modelo a ser avaliado no estudo. Nesta abordagem as possíveis alterações são tratadas como efeitos aleatórios e o modelo estatístico é feito em duas etapas.

MATERIAIS E MÉTODOS

Os dados utilizados neste trabalho são referentes a um experimento realizado na Southwest Foundation for Biomedical Research (San Antonio, TX, USA). O estudo é de interesse e tem aplicação na área de Fisiologia do Exercício, visando identificar genes DE comparados entre quatro diferentes tratamentos realizados com ratos. Considerando quatro diferentes tratamentos, foram avaliados 24 animais conforme citado:

Tratamento 1 – animais diabéticos, não exercitados (n=6 animais);

Tratamento 2 – animais diabéticos, treinados por 3 semanas, respondendo bem ao tratamento e que não eram mais diabéticos quando foram sacrificados (n=6 animais);

Tratamento 3 – animais diabéticos, treinados por 3 semanas e que não responderam bem ao tratamento e permaneceram diabéticos (n=8 animais);

Tratamento 4 – animais não diabéticos e não exercitados (n=4 animais).

Para o estudo foram utilizadas 24 lâminas de *microarrays* de cDNA, uma para cada animal e, em cada uma delas, cada tratamento foi comparado com um tratamento de referência (TR). As amostras experimentais de cada tratamento receberam fluorescência verde (Cy3) e as amostras de referência de cada tratamento receberam a fluorescência vermelha (Cy5).

No experimento, para cada animal foram avaliadas 12.768 seqüências de bases nitrogenadas, representando os genes de interesse. Estas seqüências estavam dispostas nos *spots* das lâminas em 32 blocos contendo em cada bloco 19 linhas e 21 colunas cada um deles. Desta forma, em cada *spot* é avaliado um gene de interesse (são 12.768 genes (32 blocos * 19 linhas * 21 colunas).

A medição foi feita utilizando o GenePix Pro 3.0. As fontes de fluorescência, Cy3 e Cy5, são relativamente instáveis. Por conta disso, elas podem influenciar a eficiência durante a rotulação da amostra e, assim, as intensidades detectadas por *scanners* podem ser medidas com diferentes eficiências (Speed, 2003). Baseado nestas informações, o *software* apresenta, além da medida de intensidade avaliada em cada *spot* (para cada gene), uma correção da intensidade medida também para cada *spot* (correção pelo *background*). Desta forma, para cada gene há informações da intensidade mediana para o tratamento de interesse (Cy3) e para a medida de referência (Cy5), e os respectivos valores de correção das intensidades verdes e vermelhas. Isto posto, para avaliar a intensidade da coloração verde (Cy3) ou da coloração vermelha (Cy5), deverá ser calculado o valor da intensidade de uma determinada cor menos o seu fator de correção em cada *spot*.

Nas análises de experimentos com *microarrays* há a possibilidade de encontrar valores muito altos e ordens de grandezas diferentes entre as intensidades.

No estudo serão apresentados os resultados obtidos da análise estatística realizada com os dados dos ratos submetidos aos quatro tratamentos. Serão realizadas análises visando identificar genes DE entre os tratamentos, através da ANOVA com um fator, com os dados já normalizados por um método não paramétrico. Na seqüência, será utilizada uma análise paramétrica através de um Modelo Misto seguindo o modelo de Wolfinger e colaboradores (2001) e, por último, um

modelo de ANOVA com vários fatores com efeitos fixos. Serão verificados os genes DE através destes modelos.

Modelo I – Análise dos Dados Utilizando Normalização Não Paramétrica

Neste primeiro modelo, foram feitas normalizações utilizando o método de Lowess. Com os dados já normalizados prosseguiu-se a análise, onde foi utilizado o método de Análise de variância com um fator (tratamento). A comparação entre os tratamentos foi feita para cada gene. Foram realizados ajustes dos múltiplos testes através do método de Bonferroni (*a* dividido pelo número de comparações de interesse, que para este caso foi 60.341).

A variável resposta para este modelo foi a diferença entre os logaritmos de intensidades G e R ($M_{jg} = \log_2 G_{jg} - \log_2 R_{jg}$).

Segue o modelo :

$$M_{igj} = \mu_g + T_{ig} + \varepsilon_{igj}$$

Onde:

M_{igj}: é $\log_2$ da razão das medidas das intensidade G e R normalizadas para o *j*-ésimo animal (lâmina) no *i*-ésimo tratamento do *g*-ésimo gene

g=1,...,12.768 é o número de genes avaliados no experimento

j=1,...,24 é o número de animais do experimento

i=1,...,4 é o número de tratamentos de interesse do experimento

μ_g : média global considerando o *g*-ésimo gene

T_{ig} é o efeito relativo do *i*-ésimo tratamento (comparado com o tratamento de referência)

$\varepsilon_{igj} \sim N(0, s^2)$ e independentes

Tendo M_{igj} uma distribuição Normal, com média $\mu_{ig} = \mu_g + T_{ig}$. Em termos de $\mu_{g1}, \mu_{g2}, \mu_{g3}$ e μ_{g4}, a hipótese de interesse a ser testada para este modelo é dada por:

$H_0: \mu_{g1} = \mu_{g2} = \mu_{g3} = \mu_{g4} \Leftrightarrow H_0: T_{ig} = 0 \; \forall \; i = 1, 2, 3, 4$

H_a:existe pelo menos uma diferença entre as médias

Modelo II – Anova com Modelos Mistos (Modelo de Normalização e Genético)

Este método foi realizado, considerando o modelo proposto por Wolfinger e colaboradores (2001), esta análise é realizada em dois estágios: primeiro é feita a normalização (paramétrica) dos dados. Com os resíduos deste modelo, que receberam a normalização dos dados para lâmina e tratamento, segue-se para a análise genética, ambas utilizando um modelo misto, com efeitos fixos e aleatórios. A seguir estão apresentados os modelos, considerado nas expressões I e II, em que foi utilizada como variável resposta o logaritmo (na base 2) das intensidades das medidas de expressão do *j*-ésimo animal no *i*-ésimo tratamento para o *g*-ésimo gene (y_{ijg}), sendo:

Modelo para normalização:

$$y_{ijg} = \mu + T_i + A_j + (AT)_{ij} + \varepsilon_{ijg} \text{ (expressões I)}$$

Modelo Misto Genético:

$$r_{ijg} = G_g + (GT)_{gi} + (AG)_{jg} + \gamma_{ijg} \text{ (expressões II)}$$

onde,

g=1,...,12.768 é o número de genes avaliados no experimento;

j=1,...,24 é o número de lâmina ou réplicas do experimento;

i=1,...,4 é o número de tratamentos do experimento;

No modelo genético, tendo r_{ijg} uma distribuição Normal, em termos de $\mu_{g1}, \mu_{g2}, \mu_{g3}, \mu_{g4}$ e $\mu_{g\,ref}$, que são as médias de cada tratamento para um determinado gene, a hipótese de interesse a ser testada via este modelo é dada por:

$H_0: \mu_{g1} = \mu_{g2} = \mu_{g3} = \mu_{g4} = \mu_{gref} \Leftrightarrow H_0: GT_{gi} = 0 \; \forall \; i = 1, 2, 3, 4$

H_a: existe pelo menos uma diferença entre as médias

Modelo III - Anova com Modelos de Efeitos Fixos

A seguir será aplicado o modelo sugerido por Kerr e colaboradores (2002), em que além de considerar os ajustes por gene e tratamento, o modelo considera o ajuste da lâmina, em que é realizada a normalização. O modelo é dado por:

$$y_{ijg} = \mu + A_j + T_i + (AT)_{ji} + G_g + (GT)_{gi} + (GA)_{gi} + \varepsilon_{ijg}$$

onde,

g=1,...,12768 genes avaliados no experimento

j=1,...,24 é o número de lâmina ou réplicas do experimento

i=1,...,5 é o número de tratamentos do experimento

Tendo Y_{ig} uma distribuição Normal, em termos de $\mu_{g1}, \mu_{g2}, \mu_{g3}, \mu_{g4}$ e $\mu_{g\,ref}$, que são as médias de cada tratamento para um determinado gene, a hipótese de interesse a ser testada para este modelo é dada por:

$H_0: \mu_{g1} = \mu_{g2} = \mu_{g3} = \mu_{g4} = \mu_{gref}$

H_a: pelo menos a média de um dos tratamentos não é igual

Para as análises considerando os modelos de efeitos fixos o método de estimação utilizado foi o de máxima verossimilhança. Nas análises considerando modelos de efeitos mistos o método de estimação utilizado foi o de máxima verossimilhança restrita. Nas análises foram utilizados os recursos computacionais do programa SAS 8.02 e para a análise do modelo com efeitos mistos foi utilizada a macro adaptada da versão apresentada por Wolfinger e colaboradores (2001). Os códigos da programação utilizada no ajuste dos modelos podem ser visto em Figueiredo (2006).

Em todas as análises, optou-se pelo método de Bonferroni para correção do problema dos múltiplos testes. Este método, apesar de excessivamente conservador, principalmente no contexto de *microarrays*, foi escolhido neste trabalho como um critério de correção para os múltiplos testes, visando comparar os genes DE identificados segundo diferentes abordagens de análise estatística. Ao ser adotado o método de Bonferroni, pode-se ter um aumento do número de resultados falso-negativos em prol da proteção aos erros do tipo I (resultados falso-positivos).

Comparação dos modelos

Os resultados encontrados nas análises serão comparados de duas formas, a primeira através da medida resumo dos erros-padrão de cada um dos modelos, calculando-se a média dos erros-padrão obtidos em todas as comparações. A segunda através avaliação do erro-padrão, realizando uma avaliação da eficiência relativa. Vinciotti e colaboradores (2005), sugeriu uma medida empírica para avaliar a eficiência relativa entre modelos, baseada na variância estimada para cada comparação. Esta eficiência relativa (ER) é dada por:

$$ER = \sqrt{\frac{\sum_{g=1}^{G}\sum_{z_1=1}^{Z_1} VC_{gz_1}}{\sum_{g=1}^{G}\sum_{z_2=1}^{Z_2} VR_{gz_2}}}$$

onde,

VCg_{z1} : é a variância estimada para z_1-ézima comparação do modelo de comparação no g-ésimo gene;

VRg_{z1} : é a variância estimada para z_1-ézima comparação do modelo de referência no g-ésimo gene;

$g=1,...,G$ é o número de genes.

$z_1=1,..., Z_1$ é o número de comparações realizadas no modelo de comparação;

$z_2=1,..., Z_2$ é o número de comparações realizadas no modelo de referência.

Os modelos neste trabalho não possuem a mesma quantidade de comparações, pois no Modelo I o número de níveis de tratamentos avaliados é 4, uma vez que os tratamentos de interesse são relativos ao tratamento de referência ($M_{jg} = \log_2 G_{jg} - \log_2 R_{jg}$). Nos demais, há 5 níveis de tratamentos avaliados, uma vez que o tratamento de referência é considerado como mais um nível. Desta forma optou-se por considerar as médias das variabilidades ao invés da soma, como apresentado em a seguir:

$$ER = \sqrt{\frac{(\sum_{g=1}^{G}\sum_{z_1=1}^{Z_1} VC_{gz_1}) / G * Z_1}{(\sum_{g=1}^{G}\sum_{z_2=1}^{Z_2} VR_{gz_2}) / G * Z_2}}$$

Nestas medidas de eficiência quanto menor for o valor, mais eficiente é o modelo de comparação (numerador) relativamente ao modelo de referência (denominador).

Resultados

Na Tabela 1 segue a distribuição da quantidade de genes DE para cada comparação entre tratamentos, para cada um dos modelos avaliados. Pode-se notar que cada modelo apresentou uma quantidade diferente de genes DE. Uma maior quantidade de genes foi encontrado utilizando-se o modelo 2 sugerido por Wolfinger e contribuidores (2001). Apenas os genes G1444, G5891, G5891 e G11258 foram identificados concomitante mente pelos três modelos. Entre o modelo I e o modelo II, apenas um gene foi identificado: G10690. Entre o modelo I e o modelo III, apenas dois genes foram identificados concomitantemente: G7489 e G6652. Já entre o modelo II e o modelo III 14 genes foram identificados concomitantemente. Estes dados sugerem que há muita variação nos resultados, porém uma maior semelhança entre os modelos II e III. A identificação dos genes não foi feita, pois não foi o interesse deste trabalho.

Observando os valores da Tabela 2, nota-se que, ao comparar os valores das médias dos erros padrão, o Modelo II é o que apresenta menor variabilidade, sugerindo ser o melhor modelo entre os demais, seguido do Modelo III e Modelo I, respectivamente.

Pelas comparações das eficiências relativas avaliadas na Tabela 3 em que quanto menor for o valor, mais eficiente é o modelo de comparação relativamente ao modelo de referência, nota-se que os modelo II e modelos III são mais eficiente que o modelo I. Seguindo a análise, nota-se que o modelo III é menos eficiente do que o modelo II. Desta forma, se observa os mesmos resultados da avaliação anterior, sugerindo ser o Modelo II o melhor modelo entre os demais, seguido do Modelo III e do Modelo I, respectivamente.

As comparações realizadas entre os resultados dos ajustes dos vários modelos propostos, apesar de serem de natureza descritiva, mostram que as maiores diferenças de resultados ocorreram entre os

Tabela 1 – Genes considerados DE para cada comparação entre os tratamentos para os modelos avaliados no estudo**.**

	modelo1		modelo2		modelo3	
	n	*%*	***n***	***%***	*n*	*%*
Tratamento1 x Tratamento2	4	16	5	7,1	5	11,9
Tratamento1 x Tratamento3	5	20	10	14,3	7	16,7
Tratamento1 x Tratamen to4	6	24	17	24,3	6	14,3
Tratamento2 x Tratamento3	0	0	1	1,4	1	2,4
Tratamento2 x Tratamento4	9	36	20	28,6	15	35,7
Tratamento3 x Tratamento4	1	4	17	24,3	8	19
Total	25	100	70	100	42	100

Tabela 2 – Média de Erro padrão para cada modelo considerando todas as comparações realizadas.

	Média de Erro padrão para todos os contrates
Modelo I	0,7
Modelo II	0,4
Modelo III	0,5

Tabela 3 – Eficiência relativa entre os modelos.

modelo de comparação	modelo de referência		
	Modelo I	Modelo II	Modelo III
Modelo I	1		
Modelo II	0,79	1	
Modelo III	0,85	1,07	1

procedimentos de normalização não paramétrica (Modelo I) e os demais modelos. Houve maior concordância entre os modelos de normalização paramétrica via modelos de efeitos fixos (Modelo III) e mistos em dois estágios (Modelo II). O modelo com efeitos mistos em dois estágios permitiu que um maior número de genes DE fossem identificados. Além disto, ele sugere, também, ser o melhor modelo por apresentar a menor estimativa da medida de variabilidade entre as comparações.

DISCUSSÃO

O modelo com efeitos mistos pode ser considerado um melhor modelo, pois, além de mais preciso, o modelo com efeitos mistos é mais flexível, permitindo a modelagem de covariâncias entre as observações e a inclusão de novos fatores no modelo para a normalização dos dados, além ter a possibilidade de que os erros sejam modelados com outras distribuições além do normal, como é o caso da distribuição *t*, obtendo resultados mais robustos em situações com distribuições de caudas pesadas (Lange; Little; TAYLOR, 1989).

É importante enfatizar que para a finalidade específica de comparação entre os modelos, estratégias mais sofisticadas precisam ser consideradas, como estudos de simulação e estatísticas de ajustes apropriadas ao problema. O critério de Akaike (SEARLE, 1997) não foi utilizado neste caso, dado que para a construção os modelos de normalização paramétrica seria necessário um número maior de parâmetros do que nas alternativas não paramétricas.

Por limitações como o número excessivamente grande de ajustes sob cada modelo adotado, entre outros motivos, a literatura tem recomendado que os resultados das análises estatísticas de dados de experimentos com *microarrays* sejam usados com cuidado num contexto mais exploratório e indicativo de padrão de expressão do que conclusivo.

No estudo foram avaliados três diferentes modelos: modelos com normalizações não paramétricas (Modelo I), modelos mistos em dois estágios (Modelo II) e o modelo com efeitos fixos (Modelo III). Observando os três modelos avaliados, notou-se que eles apresentam resultados diferentes. Os dois modelos que mais se aproximaram em relação aos resultados para genes DE foram os modelos II e III, o que utilizou a ANOVA da forma sugerida por Wolfinger e colaboradores (2001) com modelos mistos e o modelo de efeitos fixos sugerido por Kerr e colaboradores (2002), respectivamente. Nestes dois modelos a quantidade de genes DE foi maior do que no Modelo I, além de haver maior similaridade entre os resultados dado que genes comuns foram identificados nestes dois modelos. Em Figueiredo (2006) foi avaliado, também, o modelo semelhante ao modelo III com fatores fixos, porém em dois estágios, como sugerido no modelo II e os resultados são semelhantes aos obtidos no modelo III. O modelo II não pode ser realizado em um único estágio, devido às dificuldades computacionais para a estimação dos parâmetros.

Pelos resultados empíricos deste trabalho o modelo que conduziu a estimativas mais precisas foi o Modelo II, cujos resultados se apresentam bastante semelhante aos do modelo III. O modelo com efeitos mistos tem uma série de qualidades como, por exemplo, utilizar de forma mais apropriada as informações presentes em um experimento com *microarrays*. Ele considera como aleatórias as possíveis fontes de variação existentes nestes experimentos e, assim, leva

em consideração as variações presentes nestas fontes (lâmina, spot, etc.). É, também, um modelo, como o modelo de efeitos fixos, que permite ajuste por outras distribuições além do normal, como é o caso da distribuição *t*, na presença de distribuições assimétricas, obtendo resultados mais robustos. Uma vez que não há um interesse específico em comparar os níveis dos fatores considerados como aleatórios este modelo é vantajoso (Cui; Churchill, 2003).

CONCLUSÃO

Os experimentos com *microarrays* são bastante poderosos e de grande importância em diversas áreas. Por ser um experimento complexo e com a possibilidade de ocorrer em diversas fontes de variação, muitas técnicas de análise estatística vêm sendo estudadas visando melhores resultados com uma conclusão mais eficiente para a identificação de genes DE.

Para um melhor resultado, no início do estudo deve-se investir no bom planejamento do experimento. Devido à grande importância desta etapa da análise, diversos estudos estão sendo feitos comparando diferentes tipos de delineamentos.

Após definir um bom delineamento, a próxima etapa, que é um dos maiores esforços nestes experimentos, é minimizar a influência que fontes de variações sistemáticas podem ter nos resultados. Com este intuito há diversas alternativas de normalizações dos dados, tanto as não paramétricas, como as paramétricas que vem se apresentando como alternativas mais robustas.

Devido à grande importância e possibilidade de aplicação, um crescente interesse vem surgindo na literatura sobre os experimentos com *microarrays*. Devido às peculiaridades e dificuldades destes experimentos, as técnicas de análise estatística devem ainda ser bastante exploradas e avaliadas.

REFERÊNCIAS

CLEVELAND, W.; LOADER, C. Rejoinder to discussion of smoothing by local regression: principles and methods. In: Hardle, W.; Schimek, M. G. (Ed.) **Statistical theory and computational aspects of smoothing**. New York: Springer, 1996. p. 113-120.

CUI, X.; CHURCHILL, G. A. Statistical tests for differential expression in cDNA microarray experiments. **Genome Biol.**, London, v. 4, n. 4, p. 210, Mar. 2003

DRAGHICI, S. **Data analysis tools for DNA microarrays**. 2nd ed. Boca Raton: Chapman & Hall, 2003. 477 p.

DUDOIT, S. et al. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. ***Stat. Sinica,*** **Taiwan**, v. **2, n. 1, p.** 111-139, 2000.

FIGUEIREDO, R. A. O. **Experimentos com microarrays**: modelos para a identificação de genes diferencialmente expressos. 2006. 136 f. Dissertação (Mestrado em Estatística) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2006.

HOLDER, D. et al. Statistical analysis of high density oligonucleotide arrays: a safer approach. In: ASA Annual Meeting, 2001, Atlanta. **Proceedings**... Atlanta: [s. n.], 2001.

HUBER, W. A.; HEYDEBRECK, V.; VINGRON, M. **Handbook of statistical genetics**. 2nd ed. Chichester, NY: John Wiley, 2003.

KERR, M.; MARTIN, M.; CHURCHILL, G. A. Analysis of variance for gene expression microarray data. **J. Comput. Biol.**, New York, v. 7, n. 6, p. 819-837. 2000

KERR, M. et al. Statistical analysis of a gene expression microarray experiment with replication. *Stat. Sinica,* Taiwan, *v. 12*, n. 1, p. 203-217, jan. 2002

LANGE, K. L.; LITTLE, R. J. A.; TAYLOR, J. M. G. Robust statistical modeling using the t distribution. **J. Am. Stat. Assoc.**, New York, v. 84, n. 408, p. 881-896, dec. 1989.

ROSA, F. H. P. **Métodos adaptivos em regressão não paramétrica e normalização de microarrays de cDNA**. São Paulo: IME/USP, 2004. 21 p.

SATAGOPAN, J. M.; PANAGEAS, K. S. A statistic perspective on gene expression data analysis. **Stat. Med.**, Chichester, v. 22, n. 3, p. 481-499, feb. 2003.

SEARLE, S. R. **Linear models**. New York: Wiley-Interscience, 1997. 532 p.

SIMPSON, A. J. G.; CABALLERO, O. L. Projeto Genoma Humano e suas implicações para a saúde humana: visão geral e contribuição brasileira para o projeto. **Bioética.**, Brasília, v. 8, n. 1, p. 89-96. 2000.

SPEED, T. **Statistic analysis of gene expression microarray data**. Boca Raton: Chapman & Hall, 2003.

VINCIOTTI, V. et al. An experimental evaluation of a loop versus a reference design for two-channel microarrays. **Bioinformatics.**, Oxford, v. 21, n. 4, p. 492-501. Feb. 2005.

WOLFINGER, R. D. et al. Assessing gene significance from cDNA microarray expression data via mixed models. **J. Comput. Biol.**, New York, v. 8, n. 6, p. 625-637. 2001.

WOO, Y. et al. A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms, **J. Biomol. Tech.**, Santa Fe, v. 15, n. 4, p. 276-284. 2004.

YANG, Y. W. et al. N**ormalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation**. **Nucleic Acids Res.**, London, v. 30, n. 4, p. 15. 2002.