Modelagem de dados de Microarranjos via Teoria da resposta ao Item

Natalia Noronha Barros e **Héliton Ribeiro Tavares**

Programa de Pós-Graduação em Matemática e Estatística - PPGME, ICEN, UFPA,
Campus do Guamá, 66075110, Belém, PA
E-mail: natinobar@yahoo.com.br e heliton@ufpa.br

Introdução

A Teoria da Resposta ao Item (TRI) é uma técnica que tem sido empregada, pricipalmente em avaliações educacionais, mas seus modelos podem ser aplicados em utras áreas com as mais diversar finalidades. Com essa técnica é possível construir testes para avaliar índices sócio-econômicos (Pereira, 2004). Pode ser utilizada também, na avaliação da qualidade de vida de idosos (Guewehr, 2007), na avaliação da qualidade total (Alexandre, Andrade, Araujo & Vasconcelos, 2002), na avaliação do nível de percepção do risco sobre HIV (Ferreira, 2008) etc.

A abordagem na área da genética pode apresentar grandes avanços para ciência, pois muitas anomalias poderão ser estudadas e controladas se conecidas características associadas aos genes. Dessa maneira, este trabalho apresenta uma abordagem na área biológica, onde são ultilizado dados de expressão gênica obtidos através de estudos com microarranjos.

Objetivo

No Laboratório de Cardiologia Molecular do Instituto do Coração (INCOR-USP) estão sendo desenvolvidos estudos com linhagens de ratos congênicos objetivando identificar genes que tenham papel regulador no mecanismo da hipertensão arterial sistólica.

Um dos experimentos com microarrays foi planejado e realizado sobre um delineamento fatorial 5×2, no qual considerou-se três fatores: grupos de ratos: CH2a, CH2c, CH4, CH16 e SHR, dois níveis de sobrecarga salina: presença e ausência e cada combinação desses fatores estava presente em cinco ratos. Em cada um desses animais foram obtidos mais de 30 mil observações de fragmentos genéticos.

Partindo das considerações acima, o objetivo desse trabalho é modelar a estrutura de covariância entre os 50 tratamentos, onde cada tratamento é composto por um grupo de rato, um nível de salinidade e um animal diferente.

Metodologia

Em estudos feitos no Laboratório de Cardiologia Molecular do Instituto do Coração (INCOR-USP) foram identificadas cinco regiões cromossômicas que explicam aproximadamente 40% da pressão arterial sistólica em linhagens de ratos, indicando a presença de genes reguladores da pressão arterial nessas regiões. Nesse estudo, foram considerados dois grupos de ratos: SHR e BN, onde as regiões cromossômicas 4, 2a, 2c, 8 e 16 foram observadas. Uma maneira de verificar o efeito dessas regiões foi a realização de perturbações no sistema biológico de ratos congênicos SHR, substituindo cada uma dessas regiões pela região correspondente de um animal BN, em um total de 50 animais divididos em dois grupos, um que recebeu sobrecarga salina e o outro não. Foram obtidos os cinco grupos de animais abaixo:

CH2a : apresenta as regiões 4, 2c e 16 do SHR e a região 2a do BN.

CH2c : apresenta as regiões 4, 2a e 16 do SHR e a região 2c do BN.

CH4 : apresenta as regiões 2a, 2c e 16 do SHR e a região 4 do BN.

CH16 : apresenta as regiões 4, 2a, 2c e 16 do SHR e a região 16 do BN.

SHR : todas as regiões são do SHR.

Cada combinação entre esses grupos e um nível de salinidade estava presente em cinco ratos. Dessa maneira, cada tratamento é representado pela combinação entre um grupo, um nível de salinidade e uma repetição, a qual é composta de cinco animais distintos.

Matriz de Covariâncias

Como antecipado, nosso objetivo consiste no estudo intercorrelações entre os 50 tratamentos utilizando a modelagem de estrutura de covariâncias.

Naturalmente, as alterações genéticas que levaram à formação dos 5 grupos podem gerar resultados distintos. Os parâmetros que representarão os resultados médios de cada tratamento nos darão uma informação sobre um maior ou menor nível de desenpenho entre esses tratamentos. Embora essas informações sejam muito importantes, esses procedimentos já são bastante conhecidos na literatura, mas não na abordagem genética. No entanto, nossa busca será pelo nível de interação entre os tratamentos. Por exemplo, uma correlação alta e positiva entre dois tratamentos indica que a alteração genética que diferencia os dois tratamentos é síncrona ou inexpressiva; uma correlação alta e negativa aponta uma particular alteração genética age de forma contrária à outra alteração genética, caracterizando uma clara relação entre as funções genéticas. Uma correlação nula indica que os genes usados na alteração são claramente independentes, se forem itens distintos.

As intercorrelações entre os 50 tratamentos podem ser representadas através de uma

matriz de covariâncias de ordem 50×50 com a seguinte estrutura:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} & \boldsymbol{\Sigma}_{14} & \boldsymbol{\Sigma}_{15} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} & \boldsymbol{\Sigma}_{24} & \boldsymbol{\Sigma}_{25} \\ \boldsymbol{\Sigma}_{31} & \boldsymbol{\Sigma}_{32} & \boldsymbol{\Sigma}_{33} & \boldsymbol{\Sigma}_{34} & \boldsymbol{\Sigma}_{35} \\ \boldsymbol{\Sigma}_{41} & \boldsymbol{\Sigma}_{42} & \boldsymbol{\Sigma}_{43} & \boldsymbol{\Sigma}_{44} & \boldsymbol{\Sigma}_{45} \\ \boldsymbol{\Sigma}_{51} & \boldsymbol{\Sigma}_{52} & \boldsymbol{\Sigma}_{53} & \boldsymbol{\Sigma}_{54} & \boldsymbol{\Sigma}_{55} \end{pmatrix}.$$

As matrizes $\boldsymbol{\Sigma}_{i,j}$, para $i, j = 1, ..., 5$ correspondem às cinco repetições do experimento e cada uma delas é estruturada da seguinte maneira:

$$\boldsymbol{\Sigma}_{ij} = \begin{pmatrix} \boldsymbol{\Sigma}_{tt_{ij}} & \boldsymbol{\Sigma}_{tc_{ij}} \\ \boldsymbol{\Sigma}_{tc_{ij}} & \boldsymbol{\Sigma}_{cc_{ij}} \end{pmatrix}.$$

onde $\boldsymbol{\Sigma}_{tt_{i,j}}$, $\boldsymbol{\Sigma}_{tc_{i,j}}$ e $\boldsymbol{\Sigma}_{cc_{i,j}}$, cada uma com dimensão 5×5, contém as covariâncias entre os tratamentos nas repetições i e j, com os níveis de salinidade t e c que correspondem ao grupo que recebeu sal e grupo controle, respectivamente.

É possível que estas matrizes apresentem estruturas diferentes de acordo com o nível de salinidade, grupo e repetição. Para tornar viável o método de estimação, simplificaremos a estrutura dessa matriz para diminuir o número de parâmetros a ser estimado. A seguir mostraremos essas estruturas para esses casos:

1º caso: Covariâncias entre tratamentos com o mesmo nível de salinidade e mesma repetição: Espera-se que haja correlação entre esses tratamentos de modo que a matriz de covariâncias siga o modelo uniforme.

$$\boldsymbol{\Sigma}_{tt} = \boldsymbol{\Sigma}_{cc} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{pmatrix}.$$

2º caso: Covariâncias entre tratamentos com o mesmo nível de salinidade e repetição diferente:

$$\boldsymbol{\Sigma}^*_{tt} = \boldsymbol{\Sigma}^*_{cc} = \sigma^2 \begin{pmatrix} \rho & \rho & \rho & \rho & \rho \\ \rho & \rho & \rho & \rho & \rho \\ \rho & \rho & \rho & \rho & \rho \\ \rho & \rho & \rho & \rho & \rho \\ \rho & \rho & \rho & \rho & \rho \end{pmatrix},$$

onde ρ é a correlação entre tratamentos com mesmo nível de salinidade na mesma ou em repetições diferentes.

3º caso: Covariâncias entre tratamentos com diferentes níveis de salinidade:

$$\mathbf{\Sigma}_{tc} = \sigma^2 \begin{pmatrix} \rho_1 & 0 & 0 & 0 & 0 \\ 0 & \rho_2 & 0 & 0 & 0 \\ 0 & 0 & \rho_3 & 0 & 0 \\ 0 & 0 & 0 & \rho_4 & 0 \\ 0 & 0 & 0 & 0 & \rho_5 \end{pmatrix}.$$

Neste caso, espera-se que não haja correlação entre tratamentos com diferentes níveis de salinidade, exceto entre animais pertencentes ao mesmo grupo. Os valores de ρ_1, ρ_2, ρ_3, ρ_4 e ρ_5 representam as correlações entre tratamentos com animais nos diferentes níveis de salinidade para os grupos SHR, CH2a, Ch2c, Ch4 e CH16, respectivamente.

A metodologia para o processo de estimação será feita utilizando um modelo de probabilidade da Teoria da Resposta ao Item com uma aborgagem hierárquica bayesiana.

Teoria da Resposta ao Item

A TRI é uma técnica estatística que se fundamenta em modelos matemáticos que procuram representar a probabilidade de um indivíduo ter associado uma determinada resposta a um item em função dos parâmetros desses item e da variável latente relacionada com o respondente (Andrade, Tavares & Valle, 2000). Esses modelos dependem fundamentalmente de 3 fatores: da natureza do item, dicotômicos ou não; do número de população envolvida e da quantidade de variável latente que está sendo medida.

O modelo de Rasch ou modelo logístico de 1 parâmetro como também é chamado é um modelo unidimensional para item dicotômico. É definido por:

$$P(U_{ij} = 1|\theta_j) = \frac{1}{1 + e^{D(\theta_j - bi)}} \tag{1}$$

e adaptado à linguagem genética para os dados deste trabalho representa a probabilidade de um tratamento i ter efeito no gene j, onde

U_{ij} : é uma variável dicotômica que assume o valor 1 se o tratamento i tiver efeito no gene j e 0 se não tiver efeito;

θ_j: é a variável latente que representa a predisposição do j-ésimo gene, onde $j = 1, ..., n$, em que n é o número de genes na amostra, 35.129;

b_i : é o parâmetro de influência do tratamento i, onde $i = 1, ..., I$, em que I é o número de tratamentos, 50;

D : é o fator de escala, igual a 1.7.

com verossimilhança dada por

$$L(\boldsymbol{\zeta}) = \prod_{j=1}^{n} \prod_{i=1}^{I} P(U_{ji} = u_{ji} | \theta_j, \boldsymbol{\zeta}_i), \tag{2}$$

As intercorrelações entre os tratamentos podem estar refletidas nos parâmetros b_i através de uma matriz de covrariâncias. A estimação por máxima verossimilhança consiste apenas em encontrar estimativas para o parâmetros b_i sem levar em consideração a dependência funcional entre esses parâmetros. Dessa maneira, a metodologia para a estimação destes é a abordagem bayesiana onde considera-se que o vetor de parâmetros $\mathbf{b}$ tem como priori uma distribuição normal multivariada com vetor de médias μ e estrutura de covariâncias Σ.

Resultados

A Tabela 1 apresenta os resultados das estimativas para os parâmetros $\rho's$ aplicados aos dados.

Tabela 1: Estimativas para os parâmetros $\rho's$ com dados reais.

Parâmetro	ρ	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
Estimativa	0.704	0.889	0.901	0.623	0.290	0.599

Como mostrado anteriormente, o parâmetro ρ representa a correlação entre tratamentos com o mesmo nível de salinidade para os cinco grupos de animais. O valor de sua estimativa $\hat{\rho} = 0.704$ indica uma alta correlação positiva entre esses tratamentos, sugerindo que os grupos são bem semelhantes. Isso já era esperado, visto que, o que diferencia esses grupos são pequenas perturbações em uma dentre as cinco regiões cromossômicas que explicam aproximadamente 40% da pressão arterial nesses animais.

Os valores de ρ_1, ρ_2, ρ_3, ρ_4 e ρ_5 representam as correlações entre tratamentos com difrentes níveis de salinidade para os cinco grupos de ratos SHR, CH2a, CH2c, CH4 e CH16, respectivamente. E observando os valores das estimativas desses parâmetros, nota-se que os animais pertencentes ao grupo CH4 apresentam uma grande sensibilidade à presença do

sal, enquanto que os animais pertencentes aos grupos SHR e ao grupo CH2a foram menos sensíveis à presença do sal.

Referências

[1] Andrade, D.F., Tavares, H.R., Valle, R.C. (2000). *Teoria da Resposta ao Item: Conceitos e Aplicações*. Associação Brasileira de Estatística: São Paulo.

[2] Alexandre, J. W. C., Andrade, D.F., Araujo, A. M. S., Vasconcelos. A. P. (2002). Uma proposta de análise de um construto para medição dos fatores críticos da gestão pela qualidade através da teoria da resposta ao item: *Revista Gestão e Produção*.

[3] Ferreira, M. P., Grupo de Estudo de Pupução, Sexualidade e Aids (2008). Nível de conhecimento e percepção de risco da população brasileira sobre HIV/AIDS, 1998 e 2005: *Revista Saúde Pública 2008, 42*.

[4] Guewehr, K. (2007). *Teoria da Resposta ao Item na avaliação da qualidade de vida de idosos*. Dissertação de Mestrado, UFRGS.

[5] Pereira, V. R. (2004). *Métodos Alternativos no Critério Brasil para a Construção de Indicadores Sócio-Econômicos: Teoria da Resposta ao Item*. Dissertação de Mestrado, PUC-Rio.