INFLUÊNCIA DO NÚMERO DE OBSERVAÇÕES NO AGRUPAMENTO DE PERFIS DE EXPRESSÃO GÊNICA TEMPORAL

Roberta de Amorim Ferreira[1], Moysés Nascimento[2], Patricia Mendes dos Santos[3], Ana Carolina Campana Nascimento[4], Fabyano Fonseca e Silva[5], Laís Mayara Azevedo Barroso[6]

1. Graduanda em Matemática da Universidade Federal de Viçosa, Avenida Peter Henry Rolfs, s/n Campus Universitário 36570-000, Viçosa – MG (roberta.amorim@ufv.br)
2. Professor Doutor da Universidade Federal de Viçosa,
3. Graduanda em Matemática da Universidade Federal de Viçosa
4. Professora Doutora da Universidade Federal de Viçosa
5. Professor Doutor da Universidade Federal de Viçosa
6. Pós-Graduanda em Estatística Aplicada e Biometria da Universidade Federal de Viçosa.Brasil.

Recebido em: 30/09/2013 – Aprovado em: 08/11/2013 – Publicado em: 01/12/2013

RESUMO

A análise de dados de expressão gênica identificada ao longo do tempo – "*microarray time series*" (*MTS*) – tem possibilitado o entendimento de diversos processos biológicos uma vez que o conhecimento de grupos de genes que se expressam de forma similar possibilita inferir a respeito de funções e mecanismos reguladores desses genes. Dentre as diversas metodologias, devido ao seu apelo biológico o método proposto por RAMONI et al., (2002) apresenta destaque. Entretanto, uma característica intrínseca a estudos de *MTS* é o pequeno número de observações temporais. Assim, a utilização de métodos robustos ao pequeno número de observações temporais torna-se de extremo interesse. Assim, este trabalho teve por objetivo avaliar a robustez do método baseando na dinâmica do padrão da expressão quanto ao número de observações temporais. Foram utilizados dados que se referem à resposta de fibroblastos humanos ao soro. Estes correspondem a 517 genes cujo nível de expressão foi alterado em resposta á estimulação do soro. Esses experimentos foram repetidos sequencialmente ao longo de 12 diferentes instantes de tempo (0, 15, 30, 60, 120, 240, 360, 480,720, 960, 1200 e 1440 min.). O método avaliado é robusto, visto que o mesmo apresentou o mesmo número de grupos e altos valores de percentual de concordância até três perdas de observações temporais.

PALAVRAS-CHAVE: Fibroblastos, Expressão Gênica, séries temporais.

INFLUENCE OF THE NUMBER OF OBSERVATIONS IN CLUSTERING OF GENE EXPRESSION PROFILES

ABSTRACT

Microarray Time Series (MTS) analysis allows the researcher to characterize the gene through their longitudinal pattern of expression, since the knowledge of groups of genes that are expressed in a similar way allows the researcher to infer about the functions of genes and regulators mechanisms. Among the various methods due to its appeal biological the methodology proposed by RAMONI et al., (2002) has highlight. However, an intrinsic feature of the MTS studies is the small number of temporal observations. Thus, the use of robust methods to the small number of observations time becomes of extreme interest. This study aimed to assess the robustness of the method based on the dynamic expression pattern in the number of temporal observations. We used data relating to the response of human fibroblasts to serum. These correspond to 517 genes whose expression level changed in response to serum stimulation. These experiments were repeated sequentially over 12 different time points (0, 15, 30, 60, 120, 240, 360, 480.720, 960, 1200 and 1440 min.). The reported method is robust, given that it had the same number of groups and higher values of up to 3 percent agreement loss of temporal observations.

KEYWORDS: Fibroblast, gene expression, time series.

INTRODUÇÃO

A análise de dados de expressão gênica identificada ao longo do tempo – "*microarray time series*" (*MTS*) – tem possibilitado o entendimento de diversos processos biológicos (MUKHOPADHYAY et al., 2007), uma vez que o conhecimento de grupos de genes que se expressam de forma similar possibilita inferir a respeito de funções e mecanismos reguladores desses genes.

Devido ao grande número de genes avaliados numa análise de *MTS*, o primeiro passo para o entendimento de redes biológicas complexas é agrupar os genes que compartilham padrões similares (NASCIMENTO et al., 2012).

Dentre os métodos de análise de agrupamento utilizados, se destaca os métodos hierárquicos, o agrupamento dos genes por meio de um modelo de misturas de representações contínuas do tipo B-splines (BAR-JOSEPH et al., 2003) e os baseados na dinâmica do padrão da expressão (RAMONI et al., 2002; NASCIMENTO et al., 2012).

Dentre as metodologias citadas a metodologia proposta por e RAMONI et al. (2002) se destaca e vem sendo utilizada rotineiramente neste tipo de análise. Como por exemplo, o estudo de HEARD et al. (2006) em que se objetivou agrupar perfis de expressão de mosquitos infectados por um agente bacteriano.

Devido ao alto custo, uma característica intrínseca a estudos de *MTS* é o pequeno número de observações temporais. ERNST et al. (2005) verificaram que mais de 80% de todas as séries contidas no banco de dados de Stanford (*Stanford Microarray Database* - *SMD*) possuíam menos de oito observações temporais. Assim, a utilização de métodos robustos ao pequeno número de observações temporais torna-se de extremo interesse. Embora importante, não existe estudos

voltados para verificação da influência do número de observações no agrupamento de perfis de expressão.

Diante do exposto este trabalho teve por objetivo avaliar influência do número de observações no agrupamento de perfis de expressão por meio do método baseado na dinâmica do padrão da expressão (RAMONI et al., 2002).

MATERIAL E MÉTODOS

Para avaliar a influência do número de observações no agrupamento de perfis de expressão foram utilizados dados que se referem à resposta de fibroblastos humanos ao soro. Estes correspondem a 517 genes cujo nível de expressão foi alterado em resposta à estimulação do soro. Esses experimentos foram repetidos sequencialmente ao longo de 12 diferentes instantes de tempo (0, 15, 30, 60, 120, 240, 360, 480,720, 960, 1200 e 1440 min.). Os experimentos em questão não tinham repetições, ou seja, os valores de *fold-change* (*log*$_2$ da razão de intensidade de luz emitida pelos genes do grupo tratado e do grupo controle) são provenientes de apenas um slide de duas cores ("*two-chanel slide*") de microarranjo. Uma descrição completa dos dados pode ser encontrada no trabalho de IYER et al. (1999). Todo o conjunto de dados utilizado está disponível no seguinte endereço eletrônico: http://www.sciencemag.org/site/feature/data/984559.xhtml

O método proposto por RAMONI et al. (2002) representa as séries de expressão de acordo com equações de modelos autorregressivos e usa um procedimento bayesiano para obter a partição mais provável com base nos dados.

Considere o seguinte modelo estatístico, $Y_j = X_j\beta_j + \varepsilon_j$, em que Y_j é o vetor $\left[Y_{j(p-1)},....,Y_{jn}\right]^T, X_j$ é matriz de regressão de ordem $(n-p)\times q$ cuja t-ésima linha é representada por $\left[1, Y_{j(t-1)},....,Y_{j(t-p)}\right]$ para $t > p, q = p+1$. Os elementos do vetor $\beta_j = \left[\beta_{j0},\beta_{j1},..,\beta_{jp}\right]$ são os coeficientes autorregressão e $\varepsilon_j = \left[\varepsilon_{j(p+1)},...,\varepsilon_{jn}\right]^T$ é um vetor de erros não correlacionados que assumimos normalmente distribuído, com $E(\varepsilon_{jt}) = 0$ e variância $V(\varepsilon_{jt}) = \sigma_j^2$ para quaisquer t.

De forma sucinta, o método particiona o conjunto de perfis de expressão temporais em estudo por meio da escolha de um modelo M_c, para um conjunto de c grupos de séries temporais, constituído de c modelos autorregressivos, com coeficientes de β_k e variância σ_k^2. Cada grupo, c_k, é composto por m_k séries temporais que são modeladas conjuntamente de acordo como o seguinte modelo $Y_k = X_k\beta_k + \varepsilon_k$, em que o vetor Y_k e a matriz X_k são definidos por empilhamento dos m_k vetores y_{kj} e matrizes regressão x_{kj}, uma para cada tempo série, como segue:

$$y_k = \begin{pmatrix} y_{k1} \\ \vdots \\ y_{km_k} \end{pmatrix} \quad X_k = \begin{pmatrix} X_{k1} \\ \vdots \\ X_{km_k} \end{pmatrix}$$

O vetor ε_k é o vetor de erros não correlacionados com valor esperado zero e variância constante σ_k^2.

Dado um conjunto qualquer, formado por todas as possíveis partições, devemos classificá-las de acordo a sua probabilidade *a posteriori*. A probabilidade *a posteriori* de cada possível partição do modelo M_c é obtida por

$$P(M_c \mid y) \propto P(M_c) f(y \mid M_c)$$

em que $P(M_c)$ é a probabilidade a priori de M_c, y consiste nos dados $\{Y_k\}$, e a quantidade $f(y \mid M_c)$ é a verossimilhança marginal. A solução da verossimilhança marginal, $f(y \mid M_c)$, é dada pela integral:

$$\int f(y \mid \theta) f(\theta \mid M_c) d\theta$$

em que θ é o vetor de parâmetros especificando o agrupamento modelo M_c e $f(\theta \mid M_c)$ é distribuição a posteriori. O agrupamento bayesiano é feito por meio da escolha do modelo M_c que possui a maior probabilidade *a posteriori*.

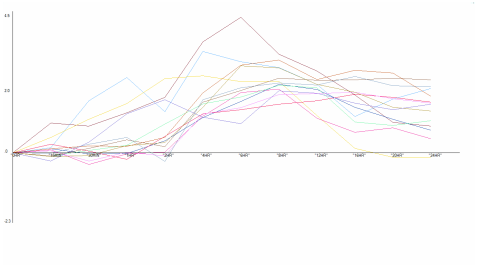
Com o objetivo de se avaliar a influência do número de observações no agrupamento de perfis de expressão por meio do método proposto por RAMONI et al. (2002) foram realizadas análises de agrupamento considerando os seguintes cenários: C1- Número total de observações temporais (12 tempos); C2- Análise considerando 11 tempos, ou seja, excluindo do conjunto de dados a última observação temporal; C3- Análise considerando 10 tempos, ou seja, excluindo do conjunto de dados as duas últimas observações temporais; C4- Análise considerando 9 tempos, ou seja, excluindo do conjunto de dados as três últimas observações temporais. Posteriormente a estas análises, calculou-se o percentual de concordância entre os grupos obtidos.

Todas as análises foram realizadas por meio do software livre CAGED (*Cluster Analysis of Gene Expression Dynamics*) que pode ser obtido gratuitamente em http://www.mybiosoftware.com/microarray-analysis/163.

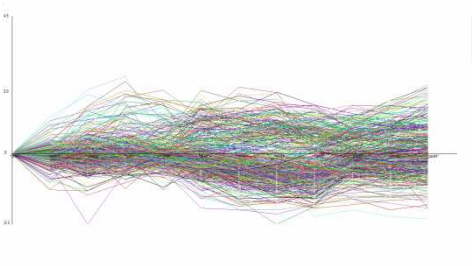
RESULTADOS E DISCUSSÃO

Após a execução do método para os quatro cenários de interesse as séries de expressão foram particionadas em 2, 2, 2 e 4 grupos para os cenários 1, 2, 3 e 4, respectivamente. O número de séries de expressão gênica em cada grupo, para cada cenário, foram 13 e 504 para o cenário 1 (Figura 1A e 1B), 9 e 509 para o cenário 2 (Figura 1C e 1D), 3 e 514 para o cenário 3 (Figura 1E e 1F) e 3, 339, 169 e 4 no cenário 4.

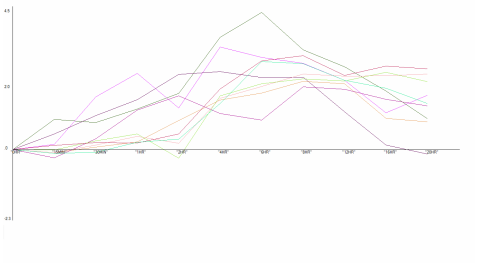
(A)



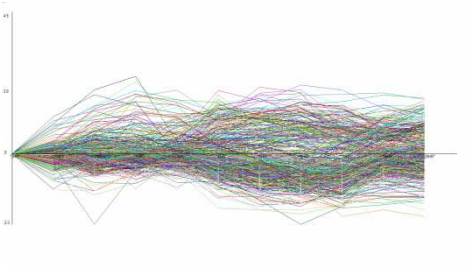
(B)



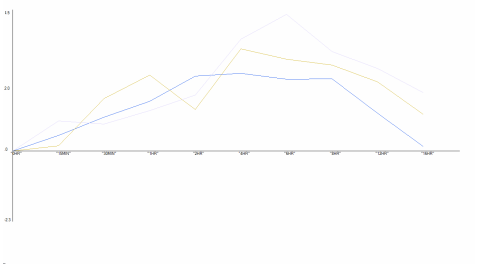
(C)



(D)



(E)



(F)

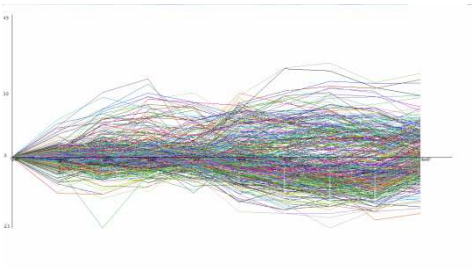


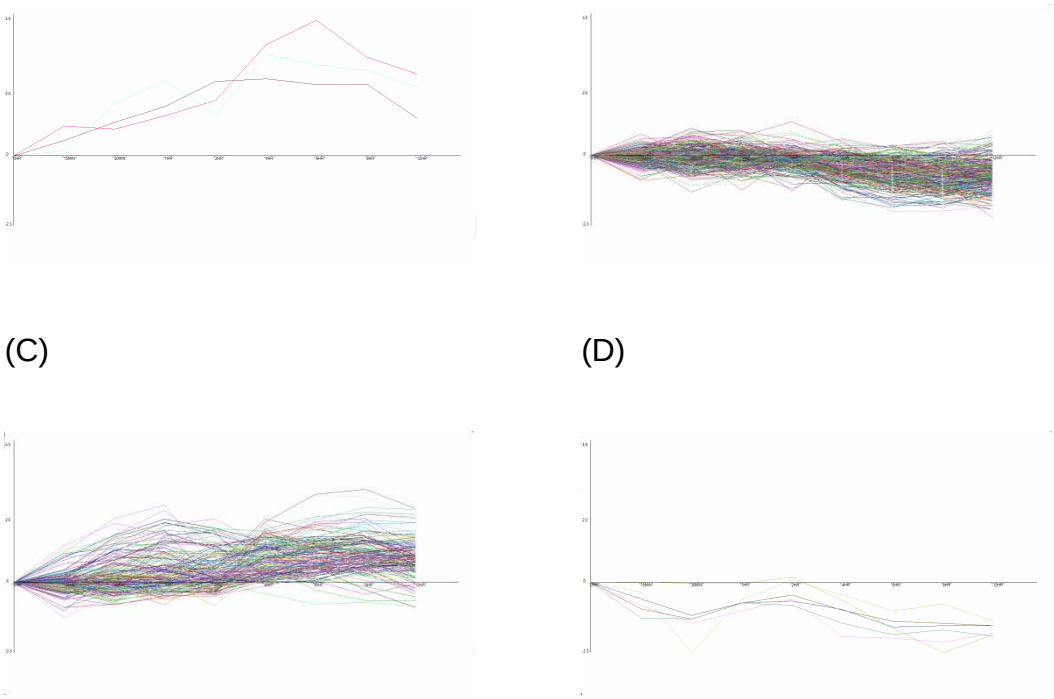
FIGURA 1. Perfis de expressão gênica: (A) grupo 1 (C1); (B) grupo 2 (C1); (C) grupo 1 (C2); (D) grupo 2 (C2); (E) grupo 1 (C3); (F) grupo 2 (C3) .
Fonte: Elaboração dos autores.

Os cenários 1 2 e 3 possuem configurações semelhantes, visto que os genes que compõem o primeiro grupo apresentam valores de expressão positivo (Figura 1). Geneticamente, essa informação é importante, pois tais genes se expressam apenas no grupo tratado (MORAIS et al. 2010). Um ponto a ser ressaltado é que o número de perfis de expressão (genes) que pertencem ao grupo 1 reduz à medida que o número de observações temporais também reduz. Desse modo percebe-se

que existe uma perda de informação quando o número de observações temporais é reduzido.

O cenário 4 apresenta uma configuração totalmente diferente das demais e assim não deve ser considerado para comparação (Figura 2).

(A) (B)



(C) (D)

FIGURA 2. Perfis de expressão gênica: (A) grupo 1 (C4); (B) grupo 2 (C4); (C) grupo 3 (C4); (D) grupo 4 (C4).

Fonte: Elaboração dos autores.

O percentual de concordância calculado, entre os cenários 1 e 2 foi de 0,99. Já em relação aos cenários 1 e 3 o percentual foi de 0,98. O percentual de concordância entre os cenários 1 e 4 não foi calculado visto que o número de grupos formados foi diferente. Esse resultado está diretamente ligado à redução no número de genes no primeiro grupo e consequentemente o aumento no segundo. Ademais, observa-se em média, 98,5% de interseção dos perfis de expressão entre os cenários. Assim, a porcentagem de se encontrar os mesmos genes no cenário original é alta, visto que, mesmo em um menor número, os perfis de expressão (genes) foram conservados.

Deve-se ficar claro que os resultados apresentados são válidos apenas para esse estudo. Em trabalhos futuros pretende-se avaliar e comparar a influência do número de repetições deste e outros métodos, tais como o não paramétrico proposto por ERNEST et al. (2005) e o apresentado por NASCIMENTO et al. (2012) em que

agrupamento é realizado por meio de uma análise conjunta bayesiana do modelo autorregressivo (AR) para dados em painel e de agrupamento, em que as estimativas dos parâmetros são consideradas variáveis de entrada no processo de agrupamento, de forma que, ao final do processo os genes que pertencem a um mesmo grupo possuem o mesmo comportamento longitudinal que é representado pelos parâmetros do modelo AR para dados em painel. Além disso, visto que no método proposto por NASCIMENTO et al. (2012) os grupos são definidos por meio de modelos AR existe a possibilidade de se obter valores para observações temporais ainda não avaliadas o que pode aumentar a robustez do método.

CONCLUSÕES

O método avaliado é pouco influenciado pelo número de observações temporais, visto que o mesmo apresentou o mesmo número de grupos e altos valores de percentual de concordância até três perdas de observações temporais.

AGRADECIMENTOS

Fundação de Amparo à Pesquisa do estado de Minas Gerais – **FAPEMIG,** pela concessão da bolsa.

REFERÊNCIAS

BAR-JOSEPH, Z.; GERBER, G. K.; GIFFORD, D. K.; JAAKKOLA, T. S.; SIMON. I. Continuous representations of time series gene expression data. **Journal of Computational Biology**, New York, v. 3, p. 341–356, 2003.

ERNST, J.; NAU, G. J.; BAR-JOSEPH, Z. Clustering short time series gene expression data. **Bioinformatics**, Oxford, v. 21, p. 159-168, 2005.

HEARD, N. A.; HOLMES, C. C.; STEPHENS, D. A.. A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves. **Journal of the American Statistical Association,** v.101, p.18, 2006.

IYER, V. R.; EISEN, M. B.; ROSS. D. T.; SCHULER, G. T.; MOORE LEE, J. C.; TRENT, J. M.; STAUDT, L. M.; HUDSON, J. JR.; BOGUSKI, M. S.; LASHKARI, D.; SHALON, D.; BOTSTEIN, D.; BROWN , P. O. The Transcriptional Program in the Response of Human Fibroblasts to Serum**, Science**, v. 283, p.83-87, 1999.

MORAIS, T. S. da S.; SILVA, F.F.; MARTINS FILHO, S.M.; SILVA, C.H.O.; NASCIMENTO, M.; SÁFADI, T. Análise Bayesiana de sensibilidade do modelo AR(1) para dados em painel: uma aplicação em dados temporais de microarrays. **Revista Brasileira de Biometria**, v.4, p171-192, 2010.

MUKHOPADHYAY, M.; CHATTERJEE, S. Causality and pathway search in microarray time series experiment. **Bioinformatics**, Oxford, v. 23, p. 442-449, 2007.

NASCIMENTO, M.; SAFADI, T.; SILVA, F. F.;NASCIMENTO, A. C. C. Bayesian model-based clustering of temporal gene expression using autoregressive panel data

approach. **Bioinformatics**, v.4, p.1-5, 2012.

RAMONI, M. F.; SEBASTIANI, P.; KOHANE, I. S. Cluster analysis of gene expression dynamics. **Proceedings of the National Academy of Sciences of the United States of America**, v.99, p.9121−9126, 2002.