Aplicação da análise de agrupamento de dados de expressão gênica temporal a dados em painel

Moysés Nascimento[1], Thelma Sáfadi[2] e Fabyano Fonseca e Silva[1]

[1]Universidade Federal de Viçosa, Departamento de Estatística, Avenida P.H. Rolfs, s/nº, CEP 36571-000 Viçosa, MG. E-mail: moysesnascim@ufv.br, fabyanofonseca@ufv.br [2]Universidade Federal de Lavras, Departamento de Ciências Exatas, Setor de Estatística e Experimentação, Campus Universitário, Caixa Postal 3037, CEP37200-000 Lavras, MG. E-mail: safadi@ufla.br

Resumo – O objetivo deste trabalho foi determinar a melhor alternativa, entre os métodos de agrupamento hierárquico (Ward) e de otimização (Tocher), para a formação de grupos homogêneos de séries de expressão gênica, e realizar previsões quanto à expressão gênica dessas séries, a partir de pequeno número de observações temporais. Os dados utilizados referem-se à expressão de genes que atuam sobre o ciclo celular de *Saccharomyces cerevisiae* e corresponderam a 114 séries de expressão gênica, cada uma com dez valores de "fold-change" (medida da expressão gênica) ao longo do tempo (0, 15, 30, 45, 60, 75, 90, 105, 120 e 135 min). As estimativas dos parâmetros dos modelos autorregressivos AR(p) foram previamente ajustadas a séries individuais (de cada gene) de dados "microarray time series" e utilizadas, como variáveis, no processo de agrupamento. As previsões da expressão gênica foram feitas dentro de cada grupo formado, a partir dos ajustes no modelo AR(p) para dados em painel. O método de Ward foi o mais apropriado para a formação de grupos de genes com séries homogêneas. Uma vez obtidos esses grupos, é possível ajustar o modelo AR(2) para dados em painel e predizer a expressão gênica em um tempo futuro (135 min), a partir de um pequeno número de observações temporais (os outros nove valores de "fold-change").

Termos para indexação: bioinformática, método de Tocher, método de Ward, microarranjo, modelo autorregressivo, série temporal.

Application of cluster analysis of temporal gene expression data to panel data

Abstract – The objective of this work was to determine the best alternative for the formation of homogeneous groups of gene expression series among the hierarchical clustering (Ward) and optimization (Tocher) methods, and to perform predictions regarding the gene expression of these series from a small number of temporal observations. The data used refer to the expression of genes that act on cell cycle of *Saccharomyces cerevisiae*, and corresponded to 114 gene expression series, with ten-fold-change values (expression measure) each, over time (0, 15, 30, 45, 60, 75, 90, 105, 120, and 135 min). The parameter estimates of autoregressive models AR(p) were previously adjusted to individual series (from each gene) of microarray time series data and used as variables in the clustering process. Gene expression predictions were made within each formed group from the adjustments in AR(p) model for panel data. The Ward's method was the more suited for the formation of gene groups with homogeneous series. Once these groups are obtained, it is possible to adjust the model AR(2) for panel-data, and successfully predict gene expression at a future time (135 min) from a small number of temporal observations (the nine other fold-change values).

Index terms: bioinformatics, Tocher's method, Ward's method, microarray, autoregressive model, 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 & Chatterjee, 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 (Costa et al., 2004).

Apesar de os métodos de agrupamento hierárquicos e os de otimização (Eisen et al., 1998) serem comumente utilizados em problemas biológicos, eles não levam em consideração a natureza sequencial das observações. Para contornar esse problema, foram desenvolvidos métodos baseados no ajuste de modelos específicos de regressão. Entre esses métodos, destacam-se os de agrupamento, que têm como base a dinâmica do padrão de expressão gênica (Ramoni et al., 2002); os modelos de Markov oculto (Schliep et al., 2003) e o agrupamento

por meio de representações contínuas do tipo B-splines (Bar-Joseph et al., 2003). Porém, apesar de úteis, eles não são adequados a experimentos relativamente pequenos ou com menos de dez observações temporais por gene (Bar-Joseph, 2004).

Ernst et al. (2005) examinaram o banco de dados de Stanford (Stanford Microarray Database, SMD) e constataram que mais de 80% de todas as séries continham menos de oito observações temporais. Esse fato evidencia a necessidade de buscar alternativas para análises de MTS, para a minimização do problema gerado pelo pequeno número de observações temporais.

Segundo Nandram & Petruccelli (1997), a utilização de modelos autorregressivos para o ajuste de dados não apenas produz previsões acuradas de valores futuros, mas também permite a análise de séries medidas em poucas observações temporais. Portanto, uma forma prática de associar métodos usuais de agrupamentos, como Ward e Tocher, às análises de dados MTS, é considerar como variáveis, na aplicação dos métodos, as estimativas de parâmetros de modelos que consideram a natureza sequencial das observações, tais como os modelos autorregressivos AR(p). Por meio dessa metodologia, é possível obter a formação de grupos de genes homogêneos quanto a suas expressões temporais.

Após a obtenção desses grupos, é possível, ainda, ajustar modelos AR(p) para dados em painel separadamente para cada grupo, o que possibilitaria o aumento da precisão das estimativas dos parâmetros em relação às análises individuais de cada série (Liu & Tiao, 1980; Silva et al., 2008; Morais et al., 2010) e, consequentemente, o aumento da precisão nas previsões de valores futuros.

A obtenção de valores da expressão gênica em tempos não estudados, proposta no presente trabalho, é uma inovação tecnológica interessante, visto que proporciona a redução de custos relacionados a procedimentos laboratoriais, que são muito elevados e podem até limitar a implantação de projetos na área de microarranjos ("microarray").

O objetivo deste trabalho foi determinar a melhor alternativa, entre os métodos de agrupamento hierárquico (Ward) e de otimização (Tocher), para a formação de grupos homogêneos de séries de expressão gênica, e realizar previsões quanto à expressão gênica dessas séries, a partir de pequeno número de observações temporais.

Material e Métodos

Os dados utilizados referem-se à expressão de genes que atuam sobre o ciclo celular de *Saccharomyces cerevisiae* (Zhu et al., 2000). Os dados originais, obtidos por meio de microarranjo de cDNA, compreendem o delineamento em comparação direta, com arranjo fatorial 2x2, em que a sincronização do ciclo celular, por meio do componente alfa (sincronizado e não sincronizado), e as diferentes cepas da levedura (selvagens e mutantes) constituíram os fatores de variação. Esses experimentos foram repetidos sequencialmente ao longo de 13 diferentes instantes de tempo equidistantes (0, 15, 30, 45, 60, 75, 90, 105, 120, 135, 150, 165 e 180 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.

Todo o conjunto de dados utilizado está disponível em: http://smd.stanford.edu/cgi-bin//publication/viewPublication.pl?pub_no=74.

Com o objetivo de ilustrar a situação descrita por Bar-Joseph (2004), ou seja, situações em que o número de observações temporais é considerado pequeno ($n \leq 10$), foram usados apenas dez tempos dos dados de células não sincronizadas; portanto, os valores de fold-change são provenientes da expressão de cepas mutantes (grupo tratado) em relação a cepas selvagens (grupo controle), dentro dessa classe de células, em cada um dos tempos avaliados.

Inicialmente, foram considerados 3.607 genes, de forma que cada um deles apresentava dez valores de fold-change. Com o intuito de realizar previsões para a décima observação, isto é, para o tempo de 135 min, consideraram-se nas análises apenas as observações referentes aos nove primeiros tempos.

As estimativas dos parâmetros, utilizadas como variáveis na análise de agrupamento, foram obtidas por meio de ajustes individuais do modelo AR(p) para cada série de expressão, ou seja, não foi considerada a teoria de dados em painel. Os modelos individuais foram ajustados tendo como variável resposta o valor de fold-change, pelo seguinte modelo:

$$Y_j = \mu + \phi_1 Y_{j-1} + \phi_2 Y_{j-2} + ... + \phi_p Y_{j-p} + \varepsilon_j,\ j = 1,2,..., n_i,$$

(modelo 1), em que: Y_j é o valor de fold-change no

tempo j; ϕ_k é o k-ésimo parâmetro de autorregressão; e ε_j é o termo de erro aleatório $\varepsilon_j \sim M(0, \sigma^2)$.

Em razão do pequeno número de observações de cada série (nove observações), ajustaram-se aos dados de MTS apenas os modelos autorregressivos de ordem $p \leq 5$. A escolha do melhor modelo foi realizada com base no critério de Schwartz, conhecido como critério de informação bayesiana (BIC). De acordo com esse critério, menores valores de BIC refletem um melhor ajuste. Sua expressão é dada por:

$$BIC(M) = nln[(L(y|\hat{\theta})] + Mln(n),$$

em que: n é o número de observações disponíveis para o ajuste; M é o número de parâmetros do modelo; e $L(y|\hat{\theta})$ é o valor assumido pela função de verossimilhança, quando se utilizam as estimativas dos parâmetros do modelo.

Com o objetivo de ilustrar a metodologia proposta, e tendo-se buscado evitar a exaustividade de informações similares, foram considerados, para a análise de agrupamento, somente genes cujo BIC indicou modelos autorregressivos de segunda ordem AR(2) como sendo os mais plausíveis e com ambos os parâmetros autorregressivos significativos ($p<0,05$). Esse critério foi escolhido tendo-se em vista a situação multiparamétrica mais simples, caracterizada pelo ajuste do modelo AR(2). As séries de expressão que não apresentaram coeficientes significativos foram descartadas, já que elas apresentaram comportamento aleatório, ou seja, os valores de expressão gênica foram independentes ao longo do tempo. É importante que o procedimento seja repetido para todos os grupos formados, ou seja, para todas as séries que apresentam o mesmo comportamento em cada ordem do modelo.

Realizou-se a análise de agrupamento tendo-se considerado como variáveis no processo de agrupamento as estimativas dos parâmetros do modelo AR(2): a média $\hat{\mu}_i$ (efeito do gene); os parâmetros de autorregressão $\hat{\phi}_{i1}$ e $\hat{\phi}_{i2}$ e a estimativa de variância do erro $\hat{\sigma}_i^2$. Assim, o agrupamento foi realizado com quatro variáveis ($\hat{\mu}$, $\hat{\phi}_1$, $\hat{\phi}_2$ e $\hat{\sigma}^2$).

Para o agrupamento das séries de expressão gênica temporal, utilizaram-se os métodos de Ward Junior (1963) e Tocher (Cruz et al., 2004; Johnson & Wichern, 2007; Ferreira, 2008). No método de Ward, os grupos são formados por meio da maximização da homogeneidade dentro do grupo, isto é, unem-se dois grupos A e B que minimizam o incremento na soma de quadrados do erro. Esse incremento é definido como $I_{AB} = n_A n_B/(n_A + n_B)\ (\bar{y}_A - \bar{y}_B)'\ (\bar{y}_A - \bar{y}_B)$, em que: n_A é o número de indivíduos pertencentes ao grupo A; n_B é o número de indivíduos pertencentes ao grupo B; $\bar{y}_A$ é o vetor de média referente aos indivíduos pertencentes ao grupo A; e $\bar{y}_B$ é a média dos valores da variável Y dentro do grupo B.

No método de Tocher, adota-se o critério de que a média das medidas de dissimilaridade, dentro de cada grupo, deve ser menor do que as distâncias médias entre quaisquer outros grupos. A entrada de um indivíduo em um grupo sempre aumenta o valor médio da distância dentro do grupo. Assim, pode-se tomar a decisão de incluir o indivíduo em um grupo, por meio da comparação entre o acréscimo no valor médio da distância dentro do grupo e um nível máximo permitido, que pode ser estabelecido arbitrariamente, ou adotado o valor máximo (θ) da medida de dissimilaridade, encontrado no conjunto das menores distâncias que envolvem cada indivíduo de seu grupo. A inclusão ou não do indivíduo k no grupo é feita, portanto, de acordo com as seguintes condições: se $d_{(grupo)k}/n \leq \theta$, inclui-se o indivíduo k no grupo; se $d_{(grupo)k}/n > \theta$, o indivíduo k não é incluído no grupo; e n representa o número de indivíduos que constitui o grupo original. Neste caso, a distância entre o indivíduo k e o grupo formado pelos indivíduos ij é dada por $d_{(ij)k} = d_{ik} + d_{jk}$ (Cruz et al., 2004). Uma característica interessante do método de Tocher é que, ao final do processo de agrupamento, o número de partições em que os indivíduos são alocados é conhecido automaticamente. Nos métodos de agrupamento hierárquicos, esse número não é fornecido, e é necessário um procedimento adicional para sua obtenção.

O número ótimo de grupos (partições), para o método de agrupamento hierárquico utilizado, foi obtido por meio do índice RMSSTD (raiz do quadrado médio do desvio-padrão), que é utilizado para calcular a homogeneidade dos agrupamentos (Khattree & Naik, 2000). O cálculo de RMSSTD, para cada novo grupo formado, foi realizado por meio da expressão,

$$RMSSTD_k = [(SQ_1 + SQ_2 + ... + SQ_p)/(gl_1 + gl_2 + ... + gl_p)]^{0,5},$$

em que:

$$SQ_j = \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2$$

representa a soma de quadrados da j-ésima variável,

calculada considerando-se as n observações presentes em cada novo grupo k, ou seja, a cada novo grupo, obtém-se um novo valor para o índice em questão; e gl_j representa o número de graus de liberdade referentes à j-ésima variável.

Para a obtenção do número ótimo de grupos, o comportamento de RMSSTD em relação ao número de grupos foi modelado exponencialmente como $RMSSTD = a(NG)^{-b}$, em que a e b são os parâmetros deste modelo e NG corresponde ao número de grupos formados. Assim, da mesma forma que no trabalho de Cecon et al. (2008) e Fiorini et al. (2010), o número ótimo de grupos foi determinado geometricamente, por meio da intersecção dessa curva com uma reta, de forma que a maior distância entre elas correspondeu ao ponto de máxima curvatura.

Após a definição do melhor método de agrupamento para a obtenção de grupos que contemplem a homogeneidade requerida para o estudo de dados em painel, realizou-se a análise tendo-se considerado a estrutura de dados em painel para cada grupo formado de acordo com o modelo,

$$Y_{ij} = \mu + \phi_{il}Y_{i(j-1)} + \phi_{i2}Y_{i(j-2)} + ... + \phi_{ip}Y_{i(j-p)} + \varepsilon_{ij},$$
$$i = 1,..., g \text{ e } j = 1, 2,..., n_i, \text{ (modelo 2)},$$

em que: Y_{ij} é o valor de fold-change do i-ésimo gene no tempo j; ϕ_{ik} é o k-ésimo parâmetro de autorregressão referente ao gene i; ε_{ij} é o termo de erro aleatório $\varepsilon_{ij} \sim N(0, \sigma^2)$.

O modelo 2 foi ajustado tendo-se considerado a técnica de variáveis indicadoras por meio do Proc Model do SAS (SAS Institute, 2009), em que é possível ajustar modelos lineares e não lineares com a estrutura de erro autorregressivo. Após o ajuste do modelo 2, obtiveram-se os valores preditos da expressão gênica em um tempo futuro (135 min).

Para avaliar a capacidade de predição do modelo 2, a última observação – valor da expressão gênica referente ao tempo 135 min – foi suprimida, para a comparação direta entre os valores preditos e os verdadeiros valores observados. Além disso, foi calculado o percentual de sinais concordantes entre os verdadeiros valores da última observação (Y_{135}) e suas estimativas ($\hat{Y}_{135}$).

Resultados e Discussão

Entre as 3.607 séries de expressão gênica, 222 apresentaram menores valores de BIC, quando modeladas por processos autorregressivos de segunda ordem, AR(2). Dessas, 114 apresentaram ambos os coeficientes autorregressivos significativos ($p<0,05$) e, portanto, foram utilizadas na análise de agrupamento.

Após a obtenção das estimativas dos parâmetros individuais de cada série, foram realizadas as análises de agrupamento pelos métodos de Tocher e Ward. Com o uso do método de Tocher, em que a medida de dissimilaridade foi o quadrado da distância euclidiana, observou-se a formação de 13 grupos, nos quais a maioria dos genes foi alocada no grupo 1 (Tabela 1).

Com o método de Ward, verificou-se que o uso de mais de 15 grupos não influencia de forma efetiva a diminuição da magnitude do RMSSTD (Figura 1 A). Esse número de grupos pode ser considerado um valor ótimo, já que, nesse ponto (Figura 1 B), a maior distância corresponde ao ponto de máxima curvatura (Cecon et al., 2008; Fiorini et al., 2010).

Para o conjunto de dados avaliados, o uso do método de agrupamento de Tocher possibilitou a formação de menor número de grupos, em que a maioria das séries de expressão gênica estava alocada no primeiro grupo (Tabelas 1 e 2). Esse fato causa menor homogeneidade das estimativas dos parâmetros μ, ϕ_1, ϕ_2 e ϕ_e^2 no primeiro grupo formado, pois seus desvios-padrão foram proporcionalmente maiores em relação aos

Tabela 1. Número de genes, médias e desvios-padrão das estimativas dos parâmetros, para cada grupo formado pelo método de Tocher, tendo-se utilizado como medida de dissimilaridade o quadrado da distância euclidiana, a partir das 114 séries de expressão gênica utilizadas no agrupamento.

Grupo	Nº genes	$\bar{\hat{\mu}}$	$\bar{\hat{\phi}}_1$	$\bar{\hat{\phi}}_2$	$\bar{\hat{\sigma}}_e^2$
1	56	0,064±0,112	0,978±0,145	-0,717±0,097	0,041±0,038
2	12	-0,106±0,122	1,266±0,073	-0,741±0,085	0,032±0,037
3	15	-0,014±0,099	0,607±0,052	-0,699±0,114	0,026±0,017
4	16	-0,072±0,097	-0,754±0,101	-0,655±0,097	0,014±0,012
5	3	0,365±0,031	0,891±0,082	-0,803±0,047	0,076±0,052
6	3	0,279±0,048	-0,805±0,178	-0,734±0,052	0,015±0,011
7	2	0,106±0,041	1,434±0,186	-0,877±0,070	0,037±0,012
8	2	-0,128±0,200	-0,917±0,088	-0,874±0,019	0,004±0,002
9	1	0,115	-1,311	-0,866	0,002
10	1	-0,021	-0,448	-0,929	0,011
11	1	-0,101	0,839	-0,918	0,061
12	1	0,573	1,003	-0,719	0,305
13	1	0,480	1,291	-0,921	0,065

$\bar{\hat{\phi}}$, média (efeito do gene); $\bar{\hat{\phi}}_1$ e $\bar{\hat{\phi}}_2$, média dos parâmetros de autorregressão; e $\bar{\hat{\sigma}}_e^2$, média da estimativa da variância do erro.

observados na maioria dos grupos formados pelo método de Ward. Essas informações indicam que o método de Ward é a alternativa de agrupamento mais interessante, quando se pretende a formação de grupos homogêneos para uso na metodologia de dados em painel, em estudos de MTS.

Do ponto de vista biológico, um maior número de grupos com menos genes, pode, consequentemente, auxiliar na descoberta de genes ligados a funções mais especificas, uma vez que elas deverão estar ligadas a um pequeno número de genes.

A Figura 2 apresenta os perfis (séries de expressão) médios de expressão gênica dos cinco primeiros grupos formados por meio do método de Ward. Verificou-se a formação de grupos de séries de expressão com comportamento diferencial, ou seja, grupos dissimilares.

Pelo princípio da análise de agrupamento, em que indivíduos semelhantes são alocados em um mesmo grupo e, consequentemente os indivíduos

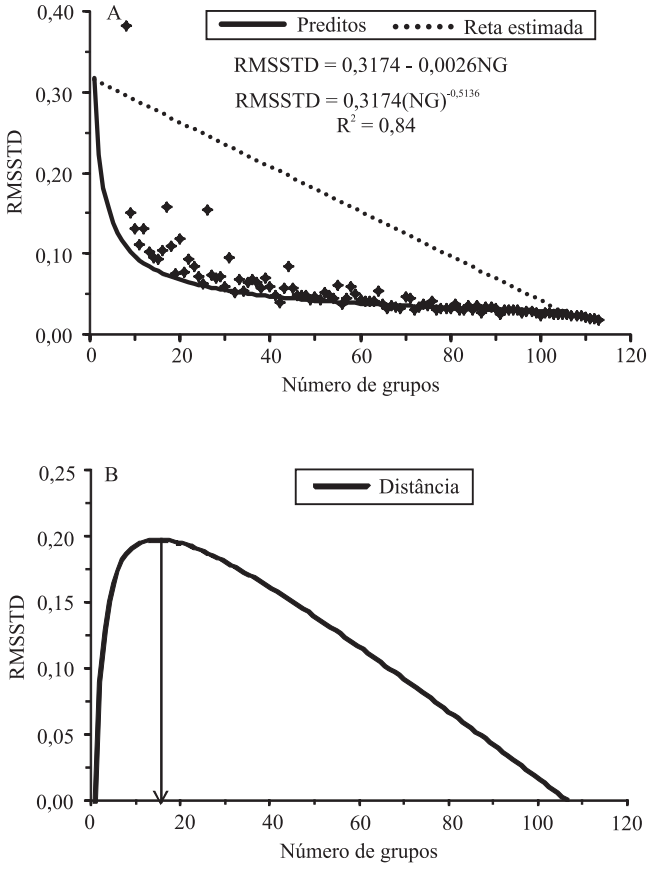

Figura 1. Determinação gráfica da influência do número de grupos formados com o uso do método de Ward, sobre a magnitude da raiz do quadrado médio do desvio-padrão (RMSSTD) (A) e do número ótimo de grupos (B), considerando-se as estimativas dos parâmetros do modelo AR(2).

Tabela 2. Número de genes, médias e desvios-padrão das estimativas dos parâmetros, para cada grupo formado pelo método de Ward, tendo-se utilizado como medida de dissimilaridade o quadrado da distância euclidiana, a partir das 114 séries de expressão gênica utilizadas no agrupamento.

Grupo	Nº de genes	$\hat{\bar{\phi}}$	$\hat{\bar{\phi}}_1$	$\hat{\bar{\phi}}_2$	$\bar{\hat{\sigma}}_e^2$
1	7	0,201±0,062	1,119±0,062	-0,843±0,041	0,038±0,014
2	16	-0,043±0,054	0,937±0,091	-0,675±0,065	0,031±0,025
3	15	0,103±0,075	0,809±0,088	-0,787±0,078	0,036±0,036
4	9	-0,003±0,077	1,227±0,049	-0,678±0,049	0,032±0,032
5	8	-0,187±0,054	1,204±0,109	-0,724±0,057	0,038±0,038
6	8	0,035±0,077	0,712±0,103	-0,573±0,023	0,074±0,070
7	5	0,316±0,073	0,883±0,060	-0,774±0,053	0,057±0,047
8	6	0,046±0,061	1,331±0,122	-0,869±0,046	0,029±0,016
9	12	-0,043±0,096	-0,801±0,102	-0,622±0,088	0,016±0,013
10	7	0,132±0,053	1,018±0,088	-0,629±0,049	0,043±0,024
11	11	-0,066±0,082	0,574±0,039	-0,768±0,090	0,023±0,018
12	7	-0,129±0,096	-0,676±0,140	-0,814±0,079	0,008±0,004
13	3	0,279±0,048	-0,805±0,178	-0,734±0,052	0,015±0,011
14	2	0,526±0,066	1,147±0,204	-0,820±0,143	0,185±0,170
15	1	0,115	-1,311	-0,866	0,002

$\hat{\bar{\phi}}$, média (efeito do gene); $\hat{\bar{\phi}}_1$ e $\hat{\bar{\phi}}_2$, média dos parâmetros de autorregressão; e $\bar{\hat{\sigma}}_e^2$, média da estimativa da variância do erro.

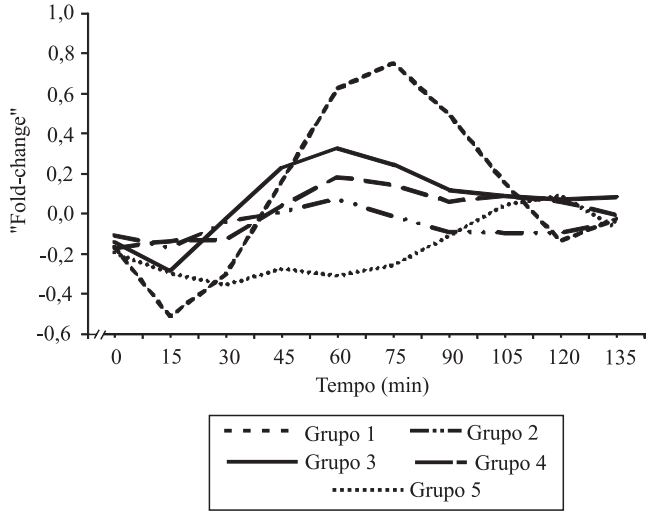

Figura 2. Perfis médios de expressão gênica dos cinco primeiros grupos formados com o uso do método de Ward, tendo-se como variáveis, no processo de agrupamento, as estimativas dos parâmetros do modelo AR(2).

que pertencem a grupos diferentes são dissimilares, o resultado verificado na Figura 2 corrobora os presentes nas Tabelas 1 e 2, ou seja, que o método de Ward é capaz de agrupar séries de comportamento similar.

O ajuste do modelo AR(2) para dados em painel, utilizado para realizar predições da expressão gênica em um tempo futuro (135 min), mostrou-se eficiente, uma vez que a percentagem de intervalos de confiança que continham os verdadeiros valores de expressão gênica, referentes ao tempo de 135 min, foi de 86%. Hay & Pettitt (2001), Silva et al. (2008) e Morais et al. (2010), que também utilizaram modelos autorregressivos para dados em painel, obtiveram valores entre 58 e 100%, o que revela que o resultado obtido no presente trabalho foi satisfatório, mesmo com o uso de apenas dez séries de expressão gênica. Esse resultado é interessante, pois os métodos de agrupamentos baseados no ajuste de modelos específicos de regressão, comumente utilizados em problemas biológicos, não são adequados (Bar-Joseph, 2004) para experimentos relativamente pequenos.

Além disso, o trabalho permite conhecer o percentual de concordância entre os sinais dos verdadeiros valores e suas estimativas, o que é uma informação interessante, pois altos valores de concordância indicam a capacidade do modelo de classificar, de forma eficiente, os genes como "up-regulated", que se expressam mais no grupo tratado, ou como "down-regulated", que se expressam mais no grupo controle. Portanto, é possível prever, em um tempo futuro, se o gene se expressará mais no tratamento ou no controle. No presente trabalho, o percentual de sinais concordantes foi de 86%, o que revela uma boa capacidade do modelo de predizer a direção do valor predito.

A obtenção de valores preditos da expressão gênica surge como uma possível forma de se reduzirem custos relacionados com os procedimentos laboratoriais para os estudos de MTS. Entretanto, deve-se ressaltar que um novo modelo pode ser ajustado para cada conjunto de dados em estudo.

Conclusões

1. A análise de agrupamento pelo método de Ward é a mais apropriada para a formação de grupos homogêneos de séries de expressão gênica.

2. O modelo para dados em painel é adequado para a predição de expressão gênica, mesmo a partir de pequenas quantidades de séries de expressão.

Referências

BAR-JOSEPH, Z. Analyzing time series gene expression data. **Bioinformatics**, v.20, p.2493-2503, 2004.

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

CECON, P.R.; SILVA, F.F. e; FERREIRA, A.; FERRÃO, R.G.; CARNEIRO, A.P.S.; DETMANN, E.; FARIA, P.N.; MORAIS, T.S. da S. Análise de medidas repetidas na avaliação de clones de café 'Conilon'. **Pesquisa Agropecuária Brasileira**, v.43, p.1171-1176, 2008.

COSTA, I.G.; CARVALHO, F. de A.T. de; SOUTO, M.C.P. de. Comparative analysis of clustering methods for gene expression time course data. **Genetics and Molecular Biology**, v.27, p.623-631, 2004.

CRUZ, C.D.; REGAZZI, A.J.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético**. 3.ed. Viçosa: UFV, 2004. 480p.

EISEN, M.B.; SPELLMAN, P.T.; BROWN, P.O.; BOTSTEIN, D. Cluster analysis and display of genome-wide expression patterns. **Proceedings of the National Academy of Sciences of the United States of America**, v.95, p.14863-14868, 1998.

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

FERREIRA, D.F. **Estatística multivariada**. Lavras: UFLA, 2008. 662p.

FIORINI, C.V.A.; SILVA, D.J.H. da; SILVA, F.F. e; MIZUBUTI, E.S.G.; ALVES, D.P.; CARDOSO, T. de S. Agrupamento de curvas de progresso de requeima em tomateiro originado de cruzamento interespecífico. **Pesquisa Agropecuária Brasileira**, v.45, p.1095-1101, 2010.

HAY, J.L.; PETTITT, A.N. Bayesian analysis of a time series of counts with covariates: an application to the control of an infectious disease. **Biostatistics**, v.2, p.433-444, 2001.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis**. 6th ed. Englewood Cliffs: Prentice Hall, 2007. 773p.

KHATTREE, R.; NAIK, D. **Multivariate data reduction and discrimination with SAS software**. Cary: SAS Institute, 2000. 574p.

LIU, L.-M.; TIAO, G.C. Random coefficient first-order autoregressive models. **Journal of Econometrics**, v.13, p.305-325, 1980.

MORAIS, T.S. da S.; SILVA, F.F. e; SILVA, C.H.O.; MARTINS FILHO, S.; 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.28, p.171-192, 2010.

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

NANDRAM, B.; PETRUCCELLI, J.D. Bayesian analysis of autoregressive time series panel data. **Journal of Business and Economic Statistics**, v.15, p.328-334, 1997.

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.

SAS INSTITUTE. **SAS/STAT**: user´s guide. Version 9.2. Cary: SAS Institute, 2009.

SCHLIEP, A.; SCHÖNHUTH, A.; STEINHOFF, C. Using hidden Markov models to analyze gene expression time course data. **Bioinformatics**, v.19, p.i264-i272, 2003.

SILVA, F.F. e; SÁFADI, T.; MUNIZ, J.A.; AQUINO, L.H. de; MOURAO, G.B. Comparação bayesiana de modelos de previsão de diferenças esperadas nas progênies no melhoramento genético de gado Nelore. **Pesquisa Agropecuária Brasileira**, v.43, p.37-45, 2008.

WARD JUNIOR, J.H. Hierarchical grouping to optimize an objective function. **Journal of the American Statistical Association**, v.58, p.236-244, 1963.

ZHU, G.; SPELLMAN, P.T.; VOLPE, T.; BROWN, P.O.; BOTSTEIN, D.; DAVIS, T.N.; FUTCHER, B. Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. **Nature**, v.406, p.90-94, 2000.

Recebido em 5 de maio de 2011 e aprovado em 2 de outubro de 2011