Análise de dados de expressão gênica: normalização de *microarrays* e modelagem de redes regulatórias

André Fujita[1], Mari C. Sogayar[2], Carlos E. Ferreira[3]

[1]Programa inter-unidades em Bioinformática da Universidade de São Paulo

[2]Instituto de Química – Universidade de São Paulo CEP 05508-900 – Cidade Universitária – SP – Brazil (Co-orientadora)

[3]Instituto de Matemática e Estatística – Universidade de São Paulo CEP 05508-090 – Cidade Universitária – SP – Brazil (Orientador)

{fujita,cef}@ime.usp.br, mcsoga@iq.usp.br

Abstract. *The goals of this work are to develop methods to analyze DNA microarray data, proposing a new normalization method, and two models to construct gene expression regulatory networks, one based on the dynamics of connectivity between genes along the cell cycle and another one which solves the dimensionality problem in which the number of microarrays experiments is smaller than the number of genes. We also present a toolbox with a user-friendly graphical interface containing several data analyses techniques and also the methods developed in this work. This work originated four papers which we published in three of the main journals of the area.*

Resumo. *Este trabalho tem como objetivos o desenvolvimento de métodos de análise de dados de microarrays, propondo uma nova forma de normalização, e dois modelos para a construção de redes regulatórias de expressão gênica, sendo uma baseada na conectividade dinâmica entre genes ao longo do ciclo celular e a outra que soluciona o problema da dimensionalidade, em que o número de experimentos de microarrays é menor que o número de genes. Apresenta-se, ainda, um pacote de ferramentas com uma interface gráfica amigável contendo diversas técnicas de análise de dados já conhecidas como também as abordagens propostas neste trabalho. Este trabalho originou quatro artigos publicados em três das principais revistas da área.*

Capítulo 1: Introdução

Nos últimos anos uma enorme massa de dados tem ficado disponível para análise de biólogos moleculares, bioquímicos e outros pesquisadores. Esta análise, entretanto, dificilmente pode ser feita manualmente, tornando imprescindível o desenvolvimento de ferramentas eficientes que auxiliem nesta análise de dados. Este é o principal objetivo deste trabalho. Em particular, estudamos três problemas de importante impacto em Biologia Computacional para os quais pudemos apresentar soluções computacionais:

1. como normalizar os dados de *microarray* a fim de permitir uma melhor análise dos resultados obtidos com esta técnica;

Endereço web da tese: http://www.teses.usp.br/teses/disponiveis/95/95131/tde-14092007-173758/

2. como inferir interações entre genes ou proteínas ao longo do ciclo celular, a fim de que biólogos possam testá-las;

3. como tratar massas de dados em que o número de observações é menor que o número de variáveis que podem influir no processo.

Apresentamos também uma ferramenta com interface amigável em que estes métodos desenvolvidos, assim como outros de uso clássico na área foram implementados,tornando-os acessíveis por pesquisadores de outras áreas.

Nos capítulos seguintes apresentamos brevemente a descrição dos métodos publicados em seus respectivos artigos.

Capítulo 2: Normalização de *microarrays*

O *microarray* é uma técnica usada para a quantificação simultânea dos níveis de expressão de milhares de genes, ou seja, para obter uma visão geral dos níveis de transcrição dos genes na célula. Um grande desafio em Bioinformática consiste no pré-processamento desses dados (normalização), ou seja, a remoção do viés existente nesta técnica.

Para a normalização de *microarrays*, propomos o uso de um método mais robusto em relação aos *outliers*, o *Support Vector Regression* (SVR) [Vapnik 1998], que pode ser útil na identificação de genes diferencialmente expressos.

Seja $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R} \times \mathbb{R}$ os dados de expressão gênica derivados dos experimentos de *microarray*, onde x é o log da intensidade de um *microarray* e y é o log da intensidade do outro. No ε-SVR [Vapnik 1998], o objetivo é obter uma função $g(x)$ que tenha no máximo ε de desvio de y_i de todos os dados, e seja o mais "plano" possível. No caso de funções lineares g:

$$g(x) = \langle w^t x \rangle + b w \in \mathbb{R}^n, b \in \mathbb{R} \text{ (1)}$$

"Planaridade" em (1) significa encontrar um w pequeno. Uma forma de assegurar isso é minimizando a norma, i.e., $\|w\|^2 = \langle w^t, w \rangle$. Isso pode ser formulado como um problema de otimização convexa:

Minimize $\frac{1}{2}\|w\|^2$ (2)

Sujeito a $y_i - \langle w^t x_i \rangle - b \leq \varepsilon \langle w^t x_i \rangle + b - y_i \leq \varepsilon$ (3)

Em (3) assumimos que existe uma função g que, com precisão ε, aproxima todos os pares (x_i, y_i), em outras palavras, assume-se que este problema de otimização é viável. Mas há casos em que isso não é possível ou queremos permitir alguns erros. Para solucionar este problema, são introduzidas as variáveis de folga ξ_i, ξ_i^* para garantir solução viável do problema de otimização, chegando a seguinte formulação [Vapnik 1998].

Minimize $\frac{1}{2}\|w\|^2 + C \sum_{i=1}(\xi_i + \xi_i^*)$ (4)

Limitado por

$$y_i - \langle w^t x_i \rangle - b \leq \varepsilon + \xi_i \text{ (5)}$$

$$\langle w^t x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \text{ (6)}$$

$$\xi_i, \xi_i^* \geq 0 \quad (7)$$

onde a constante $C > 0$ determina a quantidade acima do qual os desvios maiores que ε são tolerados.

Após obter a função $g(x)$, segue o processo de normalização cíclica descrito em [Dudoit et al. 2002].

Diversas simulações foram feitas com *microarrays* artificiais construídos conforme o modelo descrito em [Balagurunathan et al. 2002]. Em todas elas, o SVR mostrou ser mais robusto aos *outliers*, ou seja, a regressão que é a menos perturbada, tornando-se, para nossos testes, o melhor método de normalização para identificar genes diferencialmente expressos, inclusive na região de genes com alta ou baixa expressão, o que é um problema para métodos de normalização clássicos.

Este trabalho originou um artigo que recebeu mais de 2.000 acessos nos primeiros dois meses de publicação: Fujita, A., Sato, J.R., Rodrigues, L.O., Ferreira, C.E. and Sogayar, M.C. (2006) "Evaluating different methods of microarray data normalization", BMC Bioinformatics. 7:469 (*Highly accessed*).

Capítulo 3: Inferência de redes regulatórias dinâmicas

A inferência de interações em redes regulatórias é um desafio muito importante em Bioinformática e Biologia Computacional. Em particular, quando observamos um organismo ao longo do tempo, desejamos descobrir interações que são ativadas e desativadas durante o ciclo celular. Inferir tais conexões, usando o conceito de causalidade de Granger [Granger 1969] foi o objetivo deste trabalho.

Seja y_t um vetor de expressão gênica $n \times 1$ no instante t. O modelo vetor autoregressivo dinâmico (DVAR - *Dynamic Vector Autoregressive*) é definido por

$$y_t = v(t) + A_1(t)y_{t-1} + A_2(t)y_{t-2} + \cdots + A_p(t)y_{t-p} + \varepsilon_t \quad (8)$$

onde ε_t é um vetor de variáveis aleatórias com média zero e matriz de covariância $\Sigma(t)$ dado por

$$\Sigma(t) = \begin{pmatrix} \sigma_{11}^2(t) & \cdots & \sigma_{k1}(t) \\ \vdots & \ddots & \vdots \\ \sigma_{1k}(t) & \cdots & \sigma_{kk}^2(t) \end{pmatrix} \quad (9)$$

e $v(t)$ e A_i $(i = 1,2, \ldots, p)$ são respectivamente um vetor e a matriz de coeficientes, dados por:

$$v(t) = \begin{pmatrix} v_1(t) \\ \vdots \\ v_k(t) \end{pmatrix} \quad (10) \qquad A_i(t) = \begin{pmatrix} a_{11i}(t) & \cdots & a_{k1i}(t) \\ \vdots & \ddots & \vdots \\ a_{1ki}(t) & \cdots & a_{kki}(t) \end{pmatrix} \quad (11)$$

Assim, torna-se possível modelar a rede regulatória de um modo dinâmico para analisar o fluxo de informação ao longo do ciclo celular. Para estimar as funções variantes no tempo em $v(t)$, $A_i(t)$ e $\Sigma(t)$ consideramos a função de expansão *wavelet*. A *wavelet* é uma função que decompõe outras funções no domínio da freqüência, permitindo a análise em diferentes escalas de freqüência e de tempo. A idéia é que uma função $f(t)$ pode ser representada por uma combinação linear de funções *wavelet* $\psi_{j,k}(t)$, onde os índices j e k estão relacionados à escala e localização temporal,

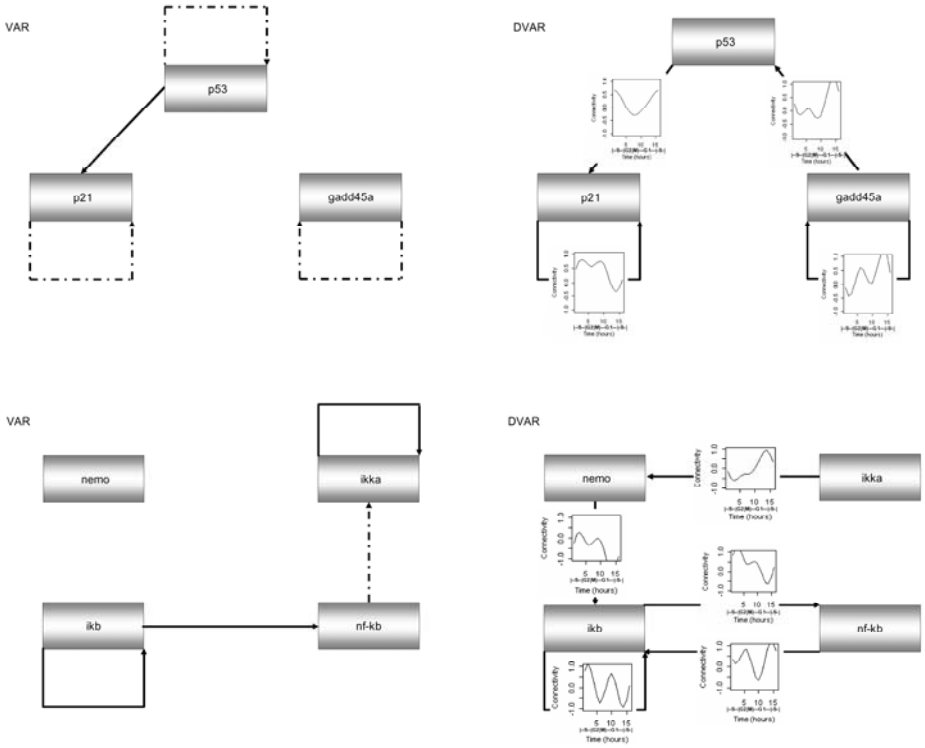
respectivamente. As funções coeficientes autoregressivos $a_{lmi}(t)$ podem ser escritas como:

$$a_{lmi}(t) = \sum_{j=k=-1}^{J} \sum_{k=0}^{2^j-1} c_{j,k}^{(i)} \psi_{j,k}(t) \quad (12)$$

onde a extensão da sére temporal T e $c_{j,k}^{(i)}$ $(j = -1,0,1,\dots,T-1; k = 0,1,\dots,2^j - 1; i = 1,2,\dots,p)$ são os coeficientes *wavelet* para o i-ésimo coeficiente da função autoregressiva $a_{lmi}(t)$ e l $(l = 1,\dots,p)$ é a ordem do vetor autoregressivo. Um ponto a ser analisado aqui é de como determinar a resolução máxima do parâmetro J. Um critério objetivo pode ser obtido por validação cruzada. Por outro lado, a escolha do grau de suavidade pode ser baseada em mudanças esperadas de acordo com informações biológicas *a priori* ou no nível de detalhe desejado. Em nossas análises, definimos J=4 porque temos a informação *a priori* de que a conectividade varia ao longo das quatro diferentes fases do ciclo celular (S, G2, M, G1).

A fim de medir o desempenho do DVAR e identificar as causalidades de Granger variantes no tempo, o DVAR foi aplicado em um conjunto de dados de células HeLa. A Figura 1 ilustra as redes estimadas usando o método proposto (DVAR) e o método clássico VAR (*Vector Autoregressive*). Nossos resultados mostram a identificação das causalidades variantes no tempo pelo DVAR.

Este trabalho originou o artigo: Fujita, A., Sato, J.R., Garay-Malpartida, H.M., Morettin, P.A., Sogayar, M.C. and Ferreira, C.E. (2007a) "Time-varying modeling of gene expression regulatory networks using the wavelet dynamic vector autoregressive method", Bioinformatics. 23:1623-1630.



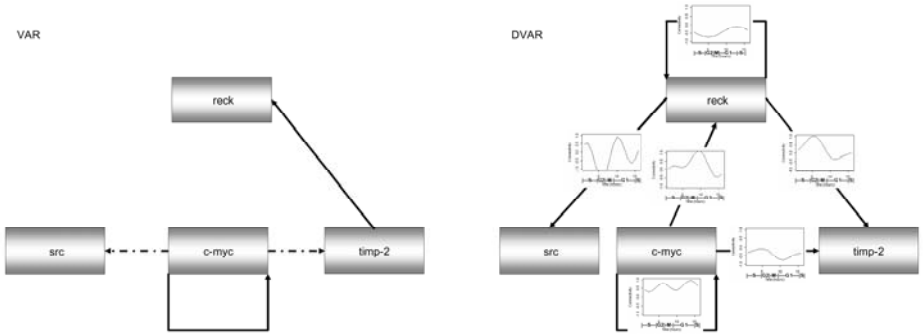


Figura 1. Redes regultórias inferidas pelos métodos VAR e DVAR. Setas pontilhadas: p < 0.10; setas contínuas: p < 0.05. Nas setas do DVAR estão representadas as causalidades de Granger variantes ao longo do ciclo celular.

Capítulo 4: Construção de redes regulatórias com número de observações inferior ao número de variáveis

Um problema que surge freqüentemente quando tratamos dados vindos de experimentos biológicos (como *microarrays*) é que o número de observações disponíveis pode ser menor que o número de variáveis que influem no processo. Inferir o funcionamento das redes sob estas condições é um problema que se coloca e exige o surgimento de novas técnicas de mineração de dados. Esta é a proposta do modelo vetor autoregressivo esparso.

Considere o modelo vetor autoregressivo esparso (SVAR - *Sparse Vector Autoregressive*) de ordem 1 como

$$y_t = A_1 y_{t-1} + \varepsilon_t \; t = 2, \ldots, T \; (13)$$

onde y_t é um vetor de expressão gênica $n \times 1$, ε_t é também um vetor $n \times 1$ com média zero e matriz de precisão Σ^{-1}, e A_1 é uma matriz $n \times n$ de parâmetros e $E(\varepsilon_t \varepsilon'_t) = \Sigma$, onde Σ é uma matriz $n \times n$ e sendo que n (número de genes) pode ser eventualmente maior que T (tamanho da série temporal). Este modelo pode ser estimado simplesmente realizando uma regressão em cada uma das variáveis nas defasagens de cada uma delas e de outras variáveis.

Assim, pode-se re-escrever o modelo como

$$Z = X\beta + E \quad E \sim N(0, \Sigma) \quad i = 1, \ldots, n \; (14)$$

onde define-se $m = T - 1$ e introduz-se a notação:

$$Z_{(m \times n)} = [y_2, \ldots, y_t, \ldots, y_T]' = [z_1, \ldots, z_i, \ldots, z_n]$$

$$\beta_{(m \times n)} = A'_1 = [\beta_1, \ldots, \beta_n]'$$

$$X_{(m \times n)} = [y_1, \ldots, y_m]'$$

$$E_{(m \times n)} = [\varepsilon_2, \ldots, \varepsilon_t, \ldots, \varepsilon_T]'$$

Por [Fan and Li 2001], [Fan and Peng 2004], [Hunter 2004] e [Hunter and Lange 2004], a regressão LASSO (*Least Absolute Shrinkage and Selection Operator*) [Tibshirani, 1996] pode ser descrita por um procedimento iterativo:

$$\hat{\beta}_i^{k+1} = (X'X + \lambda^2 D(\hat{\beta}_i^k))^{-1} X' z_i \quad i = 1, \ldots, n \text{ e } k = 1, \ldots, N_{it} \; (15)$$

onde N_{it} é o número de iterações e λ é o parâmetro que determina quanto de penalização deve ser atribuído, $D(\hat{\beta}_i^k)$ é a diagonal da matriz definida por $D(\theta) = diag(p'_\lambda(\theta)/\theta)$, com $k = 1, \ldots, n$ e $p'_\lambda(\theta) = \lambda sinal(\theta)$.

A cada iteração, os coeficientes da regressão β de cada gene são sucessivamente penalizados até atingirem o valor zero. É necessário enfatizar que o número de variáveis marcadas como zero é dependente do valor atribuído ao parâmetro λ. Então, o valor de λ foi selecionado como o valor que minimiza o valor do critério da validação cruzada generalizada (GCV - *Generalized Cross-Validation*). O valor mínimo do GCV foi atingido com o uso do algoritmo L-BFGS-B [Bryd et al. 1995].

Verificamos o desempenho do SVAR em simulações e em dados reais. O SVAR mostrou-se robusto na identificação de causalidades de Granger mesmo no contexto no qual o número de parâmetros a serem estimados é maior que o número de observações. Na aplicação em dados reais, o SVAR foi capaz de identificar conexões já conhecidas na literatura além de controlar a taxa de falsos positivos (Figura 2).

Este trabalho originou o artigo: Fujita, A., Sato, J.R., Garay-Malpartida, H.M., Yamaguchi, R., Miyano, S., Sogayar, M.C. and Ferreira, C.E. (2007b) "Modeling large gene expression regulatory networks with sparse vector autoregressive model", BMC Systems Biology. 1:39.

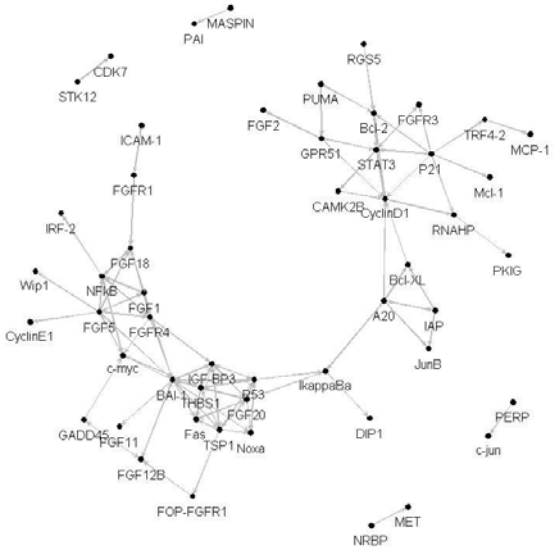


Figure 2. Rede regulatória estimada pelo método SVAR.

Capítulo 5: GEDI (*Gene Expression Data Interpreter*)

O GEDI (Figura 3) é um *software* livre sob a licensa GPL (*General Public License*) e que pode ser adquirido no *link*: http://www.iq.usp.br/wwwdocentes/mcsoga/gedi/. Este pacote possui a implementação em R (linguagem de programação estatística livre que pode ser adquirido em

http//www.r-project.org/) de todos os algoritmos apresentados nos capítulos anteriores. Este pacote de ferramentas está divido em quatro partes: (i) normalização de microarrays; (ii) identificação de genes diferencialmente expressos; (iii) classificadores de amostras e (iv) modelagem de redes regulatórias de genes.

Na seção (i) estão à disposição tanto os métodos baseados no popular método Loess, como os demais métodos mais avançados, como Splines, Wavelets e o método proposto neste trabalho, o SVR. Além destes, também adicionamos os métodos baseados em normalização por intensidade global, centralizados na média e mediana e também a normalização por quantis. Na seção (ii) tem se o clássico teste t e teste t com permutação, que permitem testar se duas médias são iguais, o teste não-paramétrico de Wilcoxon e o SAM (*Significance Analysis of Microarrays*) [Tusher and Tibshirani 2001] estão implementados. Na seção (iii) estão implementados os métodos de clusterização e classificação de amostras, baseados nos algoritmos k-means, análise discriminante linear e quadrático e o *Support Vector Machine*. Por fim, na seção (iv) estão implementados diversos métodos de modelagem de redes como o tradicional VAR, DVAR e SVAR, além das populares correlações parciais de Pearson e de Spearman.

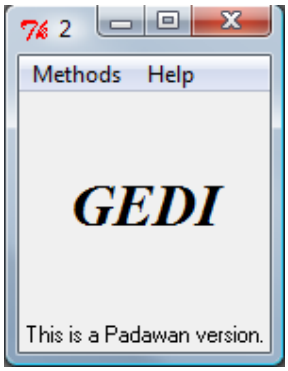


Figure 3. GEDI - *Gene Expression Data Interpreter*.

Este trabalho foi publicado recentemente e já está classificado pela revista como *Highly acessed*: Fujita, A., Sato, J.R., Ferreira, C.E. and Sogayar, M.C. (2007c) "GEDI: a user-friendly toolbox for analysis of large-scale gene expression data", BMC Bioinformatics. 8:457.

Referências

Balagurunathan, Y., Dougherty, E.R., Chen, Y., Bittner, M.L. and Trent, J.M. (2002) "Simulation of cDNA microarrays via a parameterized random signal model", Journal of Biomedical Optics. 7:507-523.

Bryd, R.H., Lu, P., Nocedal, J. and Zhu, C. (1995) "A limited memory algorithm for bound constrained optimization", SIAM J. Scientific Computing. 16:1190-1208.

Dudoit, S., Yang, Y.H., Callow, M.J., Speed, T.P. (2002) "Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments", Stat. Sin. 12:111-139.

Fan, J.Q. and Li, R.Z. (2001) "Variable selection via nonconcave penalized likelihood and its oracle properties", J. Am. Stat. Assoc. 96:1348-1360.

Fan, J.Q. and Peng, H. (2004) "Nonconcave penalized likelihood with a diverging number of parameters", Ann. Stat. 32:928-961.

Fujita, A., Sato, J.R., Rodrigues, L.O., Ferreira, C.E. and Sogayar, M.C. (2006) "Evaluating different methods of microarray data normalization", BMC Bioinformatics. 7:469 (Highly accessed).

Fujita, A., Sato, J.R., Garay-Malpartida, H.M., Morettin, P.A., Sogayar, M.C. and Ferreira, C.E. (2007a) "Time-varying modeling of gene expression regulatory networks using the wavelet dynamic vector autoregressive method", Bioinformatics. 23:1623-1630.

Fujita, A., Sato, J.R., Garay-Malpartida, H.M., Yamaguchi, R., Miyano, S., Sogayar, M.C. and Ferreira, C.E. (2007b) "Modeling large gene expression regulatory networks with sparse vector autoregressive model", BMC Systems Biology. 1:39.

Fujita, A., Sato, J.R., Ferreira, C.E. and Sogayar, M.C. (2007c) "GEDI: a user-friendly toolbox for analysis of large-scale gene expression data", BMC Bioinformatics. 8:457 (Highly accessed).

Granger, C.W.J. (1969) "Investigating causal relation by econometric and cross-sectional method", Econometrica. 37: 424-438.

Hunter, D.R. (2004) "MM algorithm for generalized Bradley-Terry models", Ann. Stat. 32:384-406.

Hunter, D.R. and Lange, K. (2004) "A tutorial on MM algorithms", Am. Stat. 58:30-37.

Tibshirani, R. (1996) "Regression shrinkage and selection via the Lasso", Journal of the Royal Statistical Society, Series B. 58:267-288.

Tusher, V. and Tibshirani, R. (2001) "Significance analysis of microarrays applied to the ionizing radiation response", PNAS. 98:5116-5121.

Vapnik, V.N. (1998) "Statistical learning theory", New York: Wiley.