Comparação de Retas de Regressão em Dados de Microarray

Cleber Martins Xavier * Gustavo H. Esteves *

Abril de 2010

Resumo Estendido

A técnica de microarray consiste da deposição de sequências de cDNA conhecidas em posições de um substrato, que pode ser uma lâmina de vidro ou uma membrana de *nylon*, que são hibridizados contra cDNAS marcados. Na utilização da membrana de *nylon*, os cDNAS são marcados radiotivamente enquanto que nas lâminas de vidro são marcados por fluorescência (canais um e dois ou *dyes cy3* ou *cy5*). Depois, essa membrana ou lâmina passa por um processo de digitalização. Nas lâminas, os fluorocromos são excitados emitindo sinais luminosos que são captados pelo *scanner*. Os dados originais são imagens que representam níveis de expressão dos genes fixados no substrato, que são analisados por um *software* específico gerando uma tabela com dados numéricos contendo os valores de intensidade de cada fragmento de DNA de interesse.

Contudo, o complexo processo experimental para a obtenção dos dados a serem analisados pode fazer com que tais valores sofram variações durante o processo de composição da tabela com os dados numéricos. Para evitar viéses que influenciam os resultados, estes dados passam por procedimentos de normalização para corrigir estes efeitos sistemáticos. Existem vários métodos capazes de resolver este problema como a normalização por energia total,

*Departamento de Estatística - Centro de Ciências e Tecnologia - Universidade Estadual da Paraíba, Campina Grande - PB, Brazil.

normalização dependente da intensidade, normalização dependente da localização espacial e ajuste de escala (ESTEVES, 2007; YANG; THORNE, 2003).

Uma vez que os dados foram normalizados, podemos seguir diversos métodos de análise, dentre eles o método para a construção de redes de relevância, que é um método de engenharia reversa (BUTTE *et al.*, 2000; BUTTE; KOHANE, 1999, 2000), para estimar perfis de interação para grupos de genes específicos. Muitas vezes são utilizados grupos gênicos associados com redes de interação gênica, ou vias metabólicas. Essas vias são formadas pelas interações de proteínas produzidas a partir de mRNAs. Essas proteínas podem interagir com o DNA regulando, tanto positiva como negativamente, a expressão de outros genes. Uma forma de se estudar redes de expressão gênica é através da avaliação das interações entre os níveis de expressão dos genes de interesse. Assim o mais interessante é tentar avaliar o perfil de interação entre os genes do grupo na tentativa de encontrar perfis de interação interessantes nos dados. Existem vários métodos matemáticos e estatísticos que tentam extrair perfis de interação entre os dados. Estes procedimentos são referenciados na literatura científica como métodos de engenharia reversa.

Muitos conjuntos de dados são planejados para comparar tipos biológicos específicos, como tecidos normais ou patológicos, por exemplo. Assim uma abordagem adequada para a avaliação de redes de expressão gênica interagentes é a construção de redes de relevância, onde as relações entre pares de genes são medidas através dos valores de correlação linear de Pearson, r. Isto é feito para cada um dos tipos de tecidos estudados, o que define para cada caso, um grafo onde os elos são conectados entre si através dos valores de r^2. Depois seleciona-se apenas os elos que apresentam valores de r^2 maiores que um certo ponto de corte $r^{2'}$, o que quebra o grafo original em sub-grafos que são chamados de redes de relevância. Esse ponto de corte é geralmente estimado através de métodos de reamostragem. Quando o interesse está em estudar as diferenças entre dois tipos de tecidos distintos, é possível buscar por pares de genes que tenham alteração significativa entre os valores de correlação dentro de cada tipo de tecido, neste caso pode-se usar a transformação Z de Fisher para avaliar a significância do resultado.

Tanto os métodos de normalização como o procedimento de construção de redes de relevância para dados de *microarray*, juntamente com vários outros métodos de análise, já estão implementados em um pacote para o ambiente de programação estatística conhecido

como R (`http://www.r-project.org`). Este pacote é intitulado *maigesPack* e está disponível para *download* através do projeto *Bioconductor* (`http://www.bioconductor.org`).

Neste trabalho pretende-se estudar a construção de redes de relevância através da utilização de modelos de regressão linear simples (FONSECA *et al.*, 1985; HOFFMANN, 2006) ao invés da correlação linear de Pearson. Ou seja, supondo que y_i seja uma variável que mede os valores de expressão de um certo gene e que x_i seja outra variável medindo os valores de expressão de um outro gene, podemos estimar a associação linear entre eles através do modelo

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \cdots, n,$$

onde n denota o número total de observações do estudo. A seguir, a idéia é utilizar um método de comparação de retas de regressão para procurar por pares de genes que tenham alteração significativa entre tipos de tecidos biológicos diferentes. Esta abordagem parece mais interessante do que a utilização do coeficiente de correlação linear de Pearson, uma vez que permite uma avaliação mais precisa da associação entre as variáveis, bem como facilita a extensão do modelo para a comparação de mais do que dois tipos de tecidos biológicos. Além disso, através da generalização do modelo simples para um modelo múltiplo, podemos buscar por perfis de associação envolvendo mais do que dois genes, o que certamente amplia o horizonte de aplicabilidade do método apresentado.

Finalmente, este método de construção de redes de relevância baseado em modelos de regressão linear simples será implementado dentro do pacote *maigesPack*, já mencionado anteriormente, e alguns conjuntos de dados de expressão gênica reais serão utilizados para a execução dos códigos e posterior comparação entre as redes geradas pelo método tradicional (usando valores de correlação linear de Pearson) e pelo novo método (usando modelos de regressão linear simples).

Referências

BUTTE, A. J.; KOHANE, I. S. Unsupervised knowledge discovery in medical databases using relevance networks. *In: Proc. AMIA Symp*, 1999.

BUTTE, A. J.; KOHANE, I. S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, v. 5,

p. 415–426, 2000.

BUTTE, A. J.; TAMAYO, P.; SLONIM, D.; GLOUB, T. R.; KOHANE, I. S. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *PNAS*, v. 97(22), p. 12182–12186, 2000.

ESTEVES, G. H. *Métodos estatísticos para a análise de dados de cDNA microarray em um ambiente computacional integrado*. Tese (Doutorado) — USP - São Paulo, 2007.

FONSECA, J. S. da; TOLEDO, G. L.; MARTINS, G. A. *Estatística aplicada*. 2. ed. [S.l.]: Atlas, 1985.

HOFFMANN, R. *Análise de Regressão: uma introdução à econometria*. 4. ed. São Paulo: Hucitec, 2006.

YANG, Y. H.; THORNE, N. P. . Normalization for two-color cdna microarray data. ims lecture notes. *Monograph Series*, v. 40, p. 403–418, 2003.