Proceedings of the 55th Session of the International Statistics Institute, 5–12 April 2005, Sydney, Australia, Paper 116.

Individual Channel Analysis of Two-Colour Microarrays

Gordon K. Smyth
Walter and Eliza Hall Institute of Medical Research, Bioinformatics
1G Royal Parade, Parkville 3050, Australia
smyth@wehi.edu.au

The traditional approach to the analysis of data from two-colour spotted microarrays is to compute the log-ratio of the expression values for each spot (Chen et al, 1997). The log-ratios are then treated as the responses in any statistical analysis of the data (Yang and Speed, 2003; Smyth, 2004). Relatively few papers have analysed spotted microarrays in terms of the separate red and green log-intensities (Kerr et al, 2000; Jin et al, 2001; Wolfinger et al, 2001). The second and third of these papers popularised a mixed model approach in which each spot is treated as a randomised block of size two.

A number of papers starting with Yang et al (2001) have summarised red and green channel intensities in terms of M-values (log-ratios) and A-values (spot log-intensities) for the purposes of graphical displays and normalisation. This paper demonstrates that the usefulness of this partition arises in good part from the fact that the M and A-values for a given spot are approximately independent even though the individual intensities are highly correlated. This paper reformulates the mixed model approach in terms of the M and A-values. This approach not only presents an efficient algorithm for estimating the mixed model but also elucidates the difference between the traditional log-ratio based approach and the analysis of individual-channels. The individual-channel approach amounts to recovering information from the between spot error stratum, i.e., from comparisons between the A-values.

There are as yet no papers which compare individual-channel with log-ratio analyses. This paper quantifies the efficiency gains which can arise from individual-channel analysis. The paper goes on to develop two new methods for individual-channel analysis which borrow information from the ensemble of probes when making inference about each individual probe. The first is an empirical Bayes method of smoothing the within and between spot components of variance. The second is based on pooling the within-spot correlation estimators. The new methods result in more stable inference than does the usual mixed model approach, especially when the number of arrays is small.

Individual channel analysis raises new and non-trivial normalisation issues in addition to those which arise in log-ratio analyses (Yang and Thorne, 2003). In this paper it will be assumed that appropriate normalisation has already been done.

M-A Models

Each spot on each microarray will yield foreground intensities estimates G_f and R_f, for the green and red channels respectively, and background estimates G_b and R_b. Write $G = \log_2(G_f - G_b)$ and $R = \log_2(R_f - R_b)$ for the green and red background corrected log-intensities. For simplicity we will assume that the background intensities have been adjusted so that all the background corrected intensities are positive and all the R and G log-intensities are defined. Write $M = R - G$ for the log-ratio and $A = (R + G)/2$ for the average intensity of the two channels.

Suppose that the experiment consists of n microarrays each printed with N genes. For gene g and array i write $y_{gi1} = G_{gi}$ for the green log-intensity and $y_{gi2} = R_{gi}$ for the red log-

intensity. Assume that $y_{gij} \sim N(\mu_{gij}, \sigma_g^2)$ where μ_{gij} is an unknown effect and σ_g^2 is an unknown variance, common across arrays and channels but possibly different across probes.

Each spot consists of a block yielding two observations. We assume that log-intensities are independent across arrays but are correlated within spots. Specifically we assume that y_{gij} and $y_{g'i'j'}$ are independent if $i \neq i'$ but that $\text{corr}(y_{gi1}, y_{gi2}) = \rho_g$ where ρ_g is an unknown intra-spot correlation. We expect ρ_g to be much larger than zero to reflect the strong correlation between intensities on the same spot. The correlations between different probes on the same microarrays, i.e., between y_{gij} and $y_{g'ij'}$ for $g \neq g'$, will be unspecified. For each spot we have $M_{gi} = y_{gi2} - y_{gi1}$ and $A_{gi} = (y_{gi2} + y_{gi1})/2$. Notice that M_{gi} and A_{gi} are independent with

$$\text{var}\, M_{gi} = \sigma_{Mg}^2 = 2\sigma_g^2(1-\rho_g), \quad \text{var}\, A_{gi} = \sigma_{Ag}^2 = \sigma_g^2(1+\rho_g)/2$$

Heteroscedastic Regression

Consider the analysis of the data y_{gij} for a given gene g. Write $\mathbf{y}_g$ for the $2n$-vector of y_{gij} and $\boldsymbol{\mu}_g$ for the $2n$-vector of μ_{gij}, $i = 1, \ldots, n$, $j = 1, 2$. The approach of Wolfinger et al (2001) is to model $\mathbf{y}_g$ with a mixed model in which the spots appear as random blocks of size two. An alternative but equivalent formulation is to represent the data in terms of the M and A-values, which converts a dependent model into an independent but heteroscedastic model. Suppose that $\boldsymbol{\mu}_g = X\boldsymbol{\beta}_g$ where X is a suitable design matrix and $\boldsymbol{\beta}_g$ is a vector of unknown coefficients. The design matrix might for example be a simple indicator matrix corresponding to different RNA sources hybridised to the arrays in which case the elements of $\boldsymbol{\beta}_g$ are mean log-intensities for those RNA sources. Write $\mathbf{M}_g = (M_{g1}, \ldots, M_{gn})^T$ and $\mathbf{A}_g = (A_{g1}, \ldots, A_{gn})^T$. Then $E(\mathbf{M}_g) = C_M^T X\boldsymbol{\beta}_g$ and $E(\mathbf{A}_g) = C_A^T X\boldsymbol{\beta}_g$ with $C_M^T = (-1, 1) \otimes I_n$ and $C_A^T = (1/2, 1/2) \otimes I_n$. The model for M and A values can be written

$$\mathbf{z}_g = Z\boldsymbol{\beta}_g + \boldsymbol{\epsilon}_g \tag{1}$$

where $\mathbf{z}_g^T = (\mathbf{M}_g, \mathbf{A}_g)^T$ is the $2n$-vector of M and A-values, $Z = (C_M, C_A)^T X$ and $\boldsymbol{\epsilon}_g$ is a vector of normal errors with diagonal covariance matrix Σ_g. The matrix Σ_g has diagonal elements equal to σ_{Mg}^2 and σ_{Ag}^2. Thus (1) defines a heteroscedastic linear model, which can be estimated using the efficient REML algorithm of Smyth (2002).

Efficiency of Individual-Channel Analysis

The formulation (1) shows that individual-channel analysis amounts to augmenting the usual log-ratio analysis by a further n-responses corresponding to the A-values. The extra information in the individual-channel analysis arises therefore from the A-values.

Replicated arrays. The simplest replicated microarray experiment consists of a series of arrays all comparing the same two RNA sources. In this case the M-values contain all the information about the fold changes and no information is gained from the A-values.

Common reference design. The second simplest design is that comparing two RNA sources, B and C say, through a common reference. Suppose that there are $n/2$ arrays comparing B with the reference and $n/2$ comparing C with the reference. Consider the analysis for a given gene. It is easily seen that the mean difference in M-values between the two groups, $\bar{M}_B - \bar{M}_C$, is an unbiased estimator the log-fold-change between the two groups with variance $8\,\sigma^2(1-\rho)/n$. The mean difference in A-values, $2(\bar{A}_B - \bar{A}_C)$, is an independent unbiased estimator with variance $8\,\sigma^2(1+\rho)/n$. The extra Fisher information provided by the A-values, as a proportion of that provided by the M-values, is therefore $(1-\rho)/(1+\rho)$. If $\rho = 0.85$, for example, the extra information provided by the A-values is about 8%.

Unconnected designs. Single channel analysis is most useful in the case of unconnected designs for which some comparisons cannot be made through the M-values. Suppose for example that $n/2$ arrays are hybridised with RNA from sources B and C and $n/2$ arrays are hybridised with sources D and E, and suppose that interest lies in the pairwise comparisons between the four RNA sources. The M-values provide information only about the direct comparisons $C-B$ and $E-D$, which are estimated with variance $4\sigma^2(1-\rho)/n$. All the other comparisons are indirect and are estimated, using both M and A-values with variance $4\sigma^2/n$. The relative efficiency of the indirect versus the direct comparisons is therefore $1-\rho$.

Small Sample Inference and Shrinkage of Variance Components

A disadvantage of the mixed model approach is that exact small sample distributions are seldom available for test statistics. The difficulty arises from the fact that the mixed model is fitted by a general residual likelihood criterion (REML) and there is no general way to associate degrees of freedom with the variance component estimators or with the standard errors of estimated coefficients or contrasts. The heteroscedastic regression model (1) gives a way around this problem. The effective degrees of freedom associated with $\hat\sigma^2_{Mg}$ and $\hat\sigma^2_{Ag}$ are d_{Mg} and d_{Ag} respectively where $d_M = n - \sum_{i=1}^n h_i$ and $d_A = n - \sum_{i=n+1}^{2n} h_i$ and the h_i are the leverages from the linear model (1). This means that approximately $\hat\sigma^2_{Mg} \sim \sigma^2_{Mg}\chi^2_{d_{Mg}}/d_{Mg}$ and $\hat\sigma^2_{Ag} \sim \sigma^2_{Ag}\chi^2_{d_{Ag}}/d_{Ag}$.

These approximations have a number of consequences. Firstly is it possible to use Satterthwaite approximations to construct approximate t-statistics for any contrasts in the model (1). Secondly it is possible to apply the empirical Bayes method of Smyth (2004) to shrink the estimators $\hat\sigma^2_{Mg}$ and $\hat\sigma^2_{Ag}$, and hence the $\hat\rho_g$, towards common values.

Common Correlation Inference

Even after shrinking, it is likely that some of the estimated within-spot correlations $\hat\rho_g$ will be negative for any given data set, an outcome which is untuitively unreasonable. One possibility is to constrain the correlations to be non-negative. Another approach is to apply a more drastic or *hard* smoothing to the correlations. The within-spot correlation arises from the technical design of two colour arrays rather than from biological variation or from characteristics of the probes or RNA targets being compared. It is therefore reasonable to assume that the correlations will be relatively consistent across the genes. This leads to the argument that the correlation estimators $\hat\rho_g$ may be pooled between genes. Under the common correlation assumption $\rho_g = \rho$, the REML estimator of $\theta = \tanh^{-1}(\rho)$ is available in closed form, $\exp(2\hat\theta) = 4(\sum_{g=1}^G \hat\sigma^2_{Ag})/(\sum_{g=1}^G \hat\sigma^2_{Mg})$. If the observations y_{gij} were independent across probes, $\hat\theta$ would be the REML estimator of θ. The estimator is generally consistent even given dependence between the probes. An alternative estimator, somewhat less efficent but also consistent, is given by

$$\frac{1}{G}\sum_{g=1}^G \{\tanh^{-1}(\hat\rho_g) - \psi(d_{Ag}/2) + \log(d_{Ag}/2) + \psi(d_{Mg}/2) - \log(d_{Mg}/2)\}$$

The pooled estimator $\hat\theta$ and the corresponding $\hat\rho = \tanh(\hat\theta)$ can be treated as known at the individual probe level, because the estimator is a consensus estimator based on all the genes. This means that the heteroscedastic regression model (1) can be transformed to an ordinary homoscedastic regression model. Simply re-scale the $\mathbf{z}_g$ to have the same variance by dividing the first n elements by $\{2(1-\hat\rho)\}^{1/2}$ and the last n elements by $\{(1+\hat\rho)/2\}^{1/2}$. The rows of Z are re-scaled the same way. This produces the ordinary regression model $\mathbf{z}_g = Z\boldsymbol\beta + \boldsymbol\epsilon_g$ with

$\boldsymbol{\epsilon}_g \sim \sigma_g^2 I_{2n}$. Methods based on this strategy has been implemented in the software package limma for R.

Data Example

The common correlation model was applied to a study by Rebecca McCracken and Steve Gerondakis at the Walter and Eliza Hall Institute involving 21 Incyte microarrays involving both direct and indirect comparisons. The median $\hat{\rho}_g$ was 0.93 after print-tip-loess normalisation of the M-values but no normalisation of the A-values, and 0.84 after quantile normalization of the A-values. This reduction in correlation shows the importance of the between-array normalization. With $\rho = 0.84$, the relative efficiency of an indirect vs a direct contrast is 0.16, i.e., direct contrasts are more than six times as efficient as indirect comparisons.

REFERENCES

Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997). Ratio based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* **2**, 364–374.

Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgel, G., and Gibson, G. (2001). The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nature Genetics* **29**, 389–395.

Kerr, M. K., Martin, M., and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**, 819–837.

Smyth, G. K. (2002). An efficient algorithm for REML in heteroscedastic regression. *Journal of Computational and Graphical Statistics* **11**, 836–847.

Smyth, G. K. (2004). Linear models and empirical Bayes for assessing differential expression in microarray experiments. *Statist. Appl. Genet. Molec. Biol.* **3**, No. 1, Article 3. **29**, 350–362.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**, 625–637.

Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. (2001). Normalization for cDNA microarray data. In *Microarrays: Optical Technologies and Informatics*, M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty (eds), Proceedings of SPIE, Vol. 4266, pages 141–152.

Yang, Y. H., and Thorne, N. P. (2003). Normalization for two-color cDNA microarray data. In: D. R. Goldstein (ed.), *Science and Statistics*, IMS Lecture Notes Volume 40, pp. 403–418.

Yang, Y. H., and Speed, T. P. (2003). Design and analysis of comparative microarray experiments. In *Statistical Analysis of Gene Expression Microarray Data*, T. P. Speed (ed.), Chapman & Hall/CRC Press, pages 35–91.

RÉSUMÉ

Individual-channel analysis of two-colour microarrays is compared with the traditional approach in terms of log-ratios. New methods are developed for individual-channel analyses using mixed models. The mixed model formulation is transformed into an "MA-Model", which is a heteroscedastic regression model for the log-ratios and log-spot-intensities. This approach (i) is the basis of a robust computational algorithm, (ii) facilitates empirical Bayes style moderation of the variance components, leading to more stable inference with small sample sizes, and (iii) provides a very simple quantification of the extra information which can be recovered from the individual-channel analysis.

L'analyse individuelle des canaux d'une biopuce (puce à ADN) à deux couleurs sera comparée avec l'approche traditionnelle utilisant des logarithmes de ratios. De nouvelles méthodes utilisant des modèles mixtes sont développes pour les analyses utilisant les canaux individuellement. La formulation du modèle mixte est transformée en un "modèle MA", qui est un modèle de régression hétéroscédastique pour les logarithmes de ratios et les logarithmes des intensités des points ("spots"). Cette approche (i) est la base d'un algorithme de calcul robuste; (ii) facilite la modération des composants de variance dans le style Bayes empirique, menant à une inférence plus stable pour des petits échantillons; et (iii) fournit une très simple quantification de l'information supplémentaire qui peut être extraite de l'analyse individuelle des canaux.