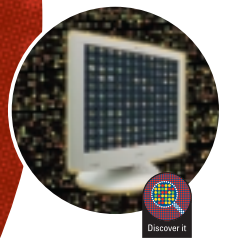
Rank Consistency Based Probe Selection for Computation of Background Adjustment and Normalization Parameters

Srinka Ghosh, Glenda Delenstarr,
Scott Connell, Paul Wolber,
Nicholas Sampas

Agilent Technologies,
3500 Deer Creek Road
Palo Alto, CA 94304



Synopsis: Microarrays are frequently hybridized with samples labeled with two or more fluorescent dyes, to compare biological sequences with controls. Data derived from such experiments are commonly contaminated with random and systematic noise. The basic data analysis approach is to quantify the former and eliminate the latter. Differential enzymatic incorporation, inherent chemical differences in dyes, gradient effects arising from hybridization and wash processes, among others, constitute potential sources of systematic variation. The elimination/minimization of the systematic error is accomplished primarily through processes of background noise removal and normalization. The efficacy of the above processes in turn depends on the selection of an unbiased and non-differentially expressed sample of probes or genes against which the statistics are computed. This poster brief describes Agilent's Rank Consistency based probe selection method and how it is employed in correcting systematic noise in microarray experiments.

Agilent Technologies

Ordering Information

www.agilent.com/chem/dna
u.s. and canada 1 800 227 9770
japan +0120 477 111
europe: marcom_center@agilent.com
global: dna_microarray@agilent.com

Printed in the U.S.A.
March 1, 2003
5988-9159EN

Agilent Technologies

ABSTRACT

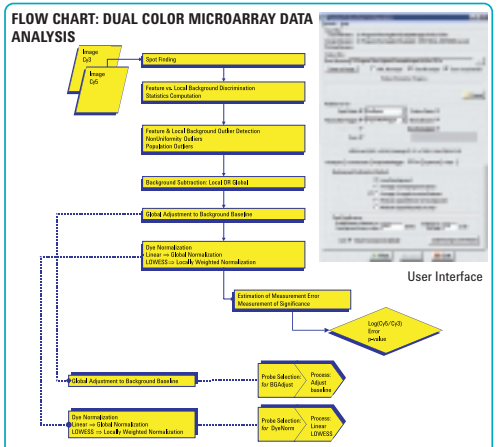
Microarrays frequently employ samples labeled with fluorescent dyes in two or more colors to compare biological sequences with a control condition. The dyes are typically incorporated into RNA targets in separate labeling reactions potentially causing differential enzymatic incorporation. Inherent chemical difference in dyes with respect to their quantum efficiencies, scanner response, and propensity for non-specific binding to array surface potentially introduce systematic inaccuracies in gene-expression measurements. Consequently, data normalization is a major Bioinformatics challenge; its success is strongly coupled to the, essential steps preceding it:

1. Background subtraction
2. Selection of probes (array features) for calculation of offset parameters to remove bias in estimated background baseline
3. Selection of probes for calculation of normalization parameters

The Rank Consistency Filter has been developed as a robust probe selection tool. Elimination of pixel and feature level outliers is a pre-condition for filtering. An initial intensity-to-rank transformation is performed, per channel, on all non-control inlier probes. The algorithm clusters those probes populating the central tendency zone (typically defined by no differential expression) of the background-subtracted intensity distribution. Probes populating the low-end of the intensity spectrum are re-clustered for background adjustment. The drivers of the algorithm are coupling of the feature intensity across the dye channels and a Euclidean distance based similarity metric.

The parameters computed using these optimized probes enhance the accuracy and reproducibility of log-ratios for two-channel data. The average standard deviation of log-ratio, across 4 self-self arrays, computed on an intra-array, inter-feature basis are 0.052 and 0.064 for only rank-consistent and all non-control probes respectively. The same metric on an inter-array, intra-feature basis yields 0.023 and 0.027 respectively.

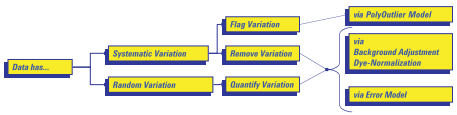
FLOW CHART: DUAL COLOR MICROARRAY DATA ANALYSIS



User Interface

All microarray data shown here, employ Agilent cDNA and in-situ arrays labeled with the fluorescent dyes - cyanine 3 and cyanine 5 to compare two labeled target samples. All images are scanned with the Agilent Microarray Scanner (G2565AA) and analyzed using Agilent Feature Extraction software G2566A, MATLAB ®, Microsoft ® Excel and Spotfire.

NEED FOR BACKGROUND ADJUSTMENT AND NORMALIZATION...1



Sources of Systematic Variation:

1. Differential enzymatic incorporation, since dyes are incorporated into RNA targets in separate labeling reactions
2. Inherent chemical differences in dyes with respect to their quantum efficiencies
3. Difference in labeling efficiency between any 2 dyes
4. Variability in amount of mRNA used for labeling, between 2 channels
5. Substrate fluorescence
6. Propensity of non-specific binding to the glass surface
7. Possible biases introduced by scanner response - Differences in the power settings of the 2 lasers
8. Gradient effects arising from hyb and wash
9. Array printing processes

Why Background subtract at all – why not work with raw data?

While the invariance of the Rank Consistency filter can be used advantageously to normalize raw signal data, most other methods of probe selection for background adjustment and normalization are sensitive to the systematic biases introduced by any or all of the above sources. Especially at the low end of the intensity distribution, the accuracy of the log ratio can be dependant on the background baseline. Since, the background levels can vary across dye channels the background subtraction needs to be performed individually in each channel on all non-outlier pixels (pixel outliers are determined via methods of inter-quartile range, standard deviation among others).
Note that the Rank Consistency filter gives robust approximations only in arrays where a statistically viable number of genes are non-differentially expressed; most biologically viable systems satisfy this criterion.

NEED FOR BACKGROUND ADJUSTMENT AND NORMALIZATION...2

Background adjustment: This mechanism corrects for inaccuracies in estimation of the noise floor in each channel; this refines feature-background discrimination especially at the lower end of the intensity spectrum.

$$\Rightarrow R(G)_{CORRECTED} = R(G)_{BGSUB} \pm R(G)_{BGADJUST}$$

For example: in case of a self versus self experiment where the feature signals should be equal in both channels across the dynamic range of the intensities, an error in background estimation in one channel over the other results in a bias. This bias is manifested as a *hook* at the low end of the signal range in case of a log ratio plot.
The background adjustment involves a two step approach:

1. Selection of probes which are subsequently used for computation and optimization of background adjustment parameters. A Rank Consistency filter coupled with intensity based thresholding filters are used as the core method.
2. The background offsets are computed in each channel from the mean/median intensities as well as dispersion of the final feature cluster from that the central tendency line.

Normalization: This is a mechanism for removing systematic variation which can potentially affect the measured level of gene expression. For example: in case of a self versus self experiment where the feature signals should be equal in both channels across the dynamic range of the intensities, differences in dye behavior can potentially cause a differential rate of change of signals.
The dye-normalization involves a two step approach:

1. Selection of probes which are subsequently used for computation and optimization of dye normalization parameters. Primary methods involve, use of (i) non-control inlier features, (ii) pre-selected house-keeping genes, (iii) virtual house-keeping genes as accomplished via the Rank Consistency filter.
   For experiments with approximately equal number of up and down regulated genes, the use of non-control inliers might suffice. But in experiments designed to probe for genes with expression, the data should be normalized using genes that are not differentially expressed and essentially populate the central tendency of the intensity distribution.
2. Normalization techniques are broadly of two types: a signal independent, non-local linear regression performed across the entire normalization data set essentially results in a global scaling of the data; a signal dependant, locally weighted linear regression method or LOWESS essentially normalizes the data taking into account the consequences of any local perturbation.

PROBE SELECTION: RANK CONSISTENCY AND PROBE SELECTION ALGORITHMS

INPUT:
BackgroundSubtracted Signal

PRIMARY FILTERS:
Non-Control Features
Non-Uniformity Inliers
Population Inliers on the basis of replicates
Saturation Inliers
Rank Consistency Inliers

RANK CONSISTENCY FILTER[1]:

$$RC_i = \frac{(|\rho_{R_i} - \rho_{G_i}|)}{N} \leq \tau$$

$Intensity_{RED}(I_R) \rightarrow Rank_{RED}(\rho_R)$;
$Intensity_{GREEN}(I_G) \rightarrow Rank_{GREEN}(\rho_G)$;
N: Total number of features on array; i is the feature index
τ: Rank Consistency threshold; it is possible that τ is a constant ⇒ τ = τ_c,
OR τ is a function of intensity ⇒ τ = τ(Intensity)
F1: Probes used for computation of normalization parameters
F2: Probes used for computation of global background adjustment offsets

FIGURE 1: Features Passing the Rank Consistency Filter: F1(blue)

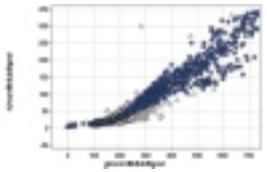


FIGURE 2: Features Passing Low End Filter cutoff(C1)(blue)

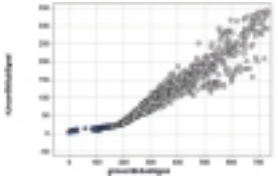


FIGURE 3: Features Populating a Thresholded Dispersion Envelope(TDE) (blue)

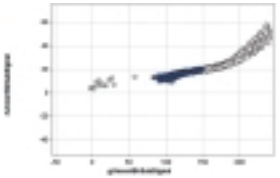


FIGURE 4: Additional Control Features Populating the Region defined by TDE(green)

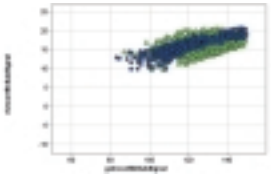
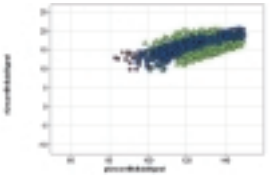


FIGURE 5: Features Passing Final Low End Filter cutoff(C2): F2 (pink)



PROBE SELECTION: PERFORMANCE OF RANK CONSISTENCY FILTER

The two other commonly used probe selection methods are :

1) All Significant, Non-control, Non-Outlier Features
2) Pre-selected HouseKeeping Genes

Ideally, the probes used for determination of normalization parameters should themselves exhibit no differential expression. As shown in the adjacent data, normalization probes selected via method (1) (figure 6b) encompass regions outside the central tendency of the data in comparison to Rank Consistency (figure 6a). Table 1, compares the two methods on both intra and inter array as well inter and intra feature basis. For performance evaluation, the data analysis was performed on 4 self versus self arrays where the log-ratios were computed from background subtracted signal that had been background adjusted and dye-normalized.

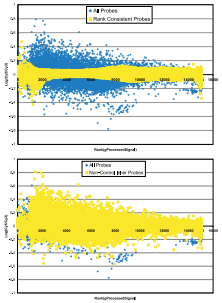


FIGURE 6: Comparison of Probe Selection Methods:
(a)Rank Consistency[9491/15185] versus (b)Non-Control Inliers 14170/15185]

The standard deviation measurements on the Log(Cy5/Cy3), demonstrate a tighter dispersion in case of the use of Rank Consistency probes. Coefficient of Variability(CV) evaluated across replicate probes both on the same and across arrays basis can be implemented as another metric for characterization of the two approaches. A similar analysis can be used to compare the efficacy of the pre-selected HouseKeeping genes with that of the virtual selection process. The inter-array, intra-feature average standard deviation metric can also be applied to arrays demonstrating differential expression.

TABLE 1: Comparison of Probe Selection Methods:
Rank Consistency versus Non-Control Inliers on intra-array, inter-feature and inter-array, intra-feature basis; N = # of arrays = 4 self versus self;

Average Standard Deviation(Log(Cy5/Cy3))	Rank Consistent Probes	Non-Control Inlier Probes
Intra-array[Inter-feature]	0.052	0.064
Inter-array[Intra-feature]	0.023	0.027

PROCESS: IMPACT OF BACKGROUND ADJUSTMENT AND DYE-NORMALIZATION

Data from a self versus self array has been used to illustrate the impact of background adjustment and dye-normalization in a step-by-step manner. In the adjoining figures, all features on the array are shown, with the rank consistent ones highlighted in blue.
Figure 7 highlights the dispersion in the data introduced by the systematic and random sources of error. The low-end hook is primarily an artifact of mis-estimation of background signal in one channel over the other. The high end distortion is primarily due the dye bias across the two channels. The effect of this curvature is also manifested in a Log(Cy3) versus Log(Cy5) plot.
Figure 8, elucidates the effect of the global background adjustment as applied to the back ground subtracted signals from both channels.The background adjustment algorithm employed here allows computation of offsets in each dye channel where the effect of each channel is given equal weight. In case of use of a local background subtraction method, the coupling with a global background adjustment method facilitates a hybrid approach to background correction and removes the hook artifact.
Figure 9, elucidates the impact of LOWESS dye-normalization on the background adjusted data. The symmetrization of the log ratio with respect to the line of no differential expression - Log(Cy5/Cy3) = 0 - provides a visual validation of the combined techniques.

FIGURE 7: Log(Cy5/Cy3) where log ratio is computed on background subtracted signal only. The data shows curvatures at both low and high ends. Data obtained from self versus self array.

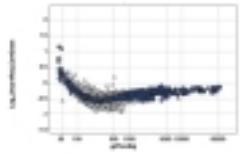


FIGURE 8: Log(Cy5/Cy3) following global background adjustment only.

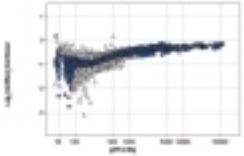
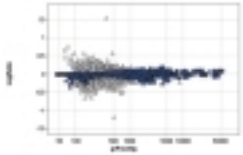


FIGURE 9: Log(Cy5/Cy3) following global background adjustment and dye-normalization corrections.



SUMMARY AND CONCLUSIONS

Probe Selection:

1. The Rank Consistency estimates the envelope for central tendency of the gene expression data. Thereby robustly identifying genes/probes which express no differential expression;
2. The Rank Consistency method provides a robust mechanism for probe selection; these probes are subsequently used for computation and optimization of background adjustment and dye-normalization;
3. The efficacy of the filter can be optimized via a method of pre-qualification of probes, as shown by the exclusion of intensity non-uniformity, saturation and populations outliers;

Process:

1. Local background subtraction coupled with global background adjustment provides a hybrid background estimation technique.
2. Background adjustment and dye-normalization processes are instrumental in reduction/elimination of systematic variation in a biological dataset.

Finally:

For a given set of experimental conditions, the efficacy of the Rank Consistency filter can be compared against other probe selection methods via the implementation of a dispersion metric on Log(Cy5/Cy3).

ACKNOWLEDGEMENTS

The authors would like to thank the following people from Agilent Bio Research Solutions and Labs for their insights and thoughts: Mel Kronick, Karen Shannon, Bo Curry.

1. *Issue in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects,* Wing Hung Wong et al, Nucleic Acids Research, 2001, Vol 29, No 12, pp 2549-2557.