

SUPPLEMENTAL DATA: ROBUST ESTIMATORS FOR EXPRESSION ANALYSIS

EARL HUBBELL, WEI-MIN LIU, AND RUI MEI

ABSTRACT. This is supplemental data extracted from the paper *Robust Estimators for Expression Analysis* and provided on this web-site.

1. ON PROPERTIES OF THE SIGNAL ALGORITHM

One advantage of using a logarithmic transformation is that it approximately stabilizes the variance of the resulting estimate. This stabilization can be observed in figure 1. Note that as expected, this approximation breaks down at the low end where the true concentration approaches zero.

One nice property of the biweight as opposed to pure outlier rejection is that down-weighting of extreme values is done in a smooth fashion. Figure 2 illustrates this phenomenon on real data.

2. ON IMPUTATION

While a superficially attractive idea, dropping censored values where $I_{MM} > I_{PM}$ reduces the number of observations (conceivably halving the effective number of probe pairs by chance near zero concentration) and has the undesired result of increasing the variance and decreasing the ability to resist outliers. This can lead to an increase in the observed variance due to probe affinity effects. Therefore, we use an imputation procedure in preference to dropping such observations.

As an example, consider figure 3. This figure shows observed probe intensities for three spike concentrations, at 0, 1, and 2 pM. At the zero concentration spike, the observed typical proportion of mismatch to perfect match is 97% over the whole probe set, 76% for 1 pM, and 68% for 2 pM. It is clear that some mismatch values (shown in grey) (for example, probe pairs 2 and 13) cannot represent the stray signal for their corresponding perfect match intensity (shown in black). (Note that their respective perfect match probes intensities increase with concentration, indicating that the perfect match value is tracking the target). For such cases, we substitute an estimate of stray signal (shown in white) that has the typical ratio to the observed perfect match value. Note that this estimate (which is computed as a proportion) takes on different absolute values for different perfect match intensities.

3. ON THE LATIN SQUARE DATA SETS

Thirteen concentration levels were chosen to be evenly spaced on a logarithmic scale, ranging from .25 pm to 1024 pm by doubling. In addition, the fourteenth concentration level was zero concentration. Fourteen human transcripts were chosen for hybridization in 42 experiments with three-fold replication of a 14x14 cyclic latin square. Two of the

Key words and phrases. Expression analysis, robust estimators.

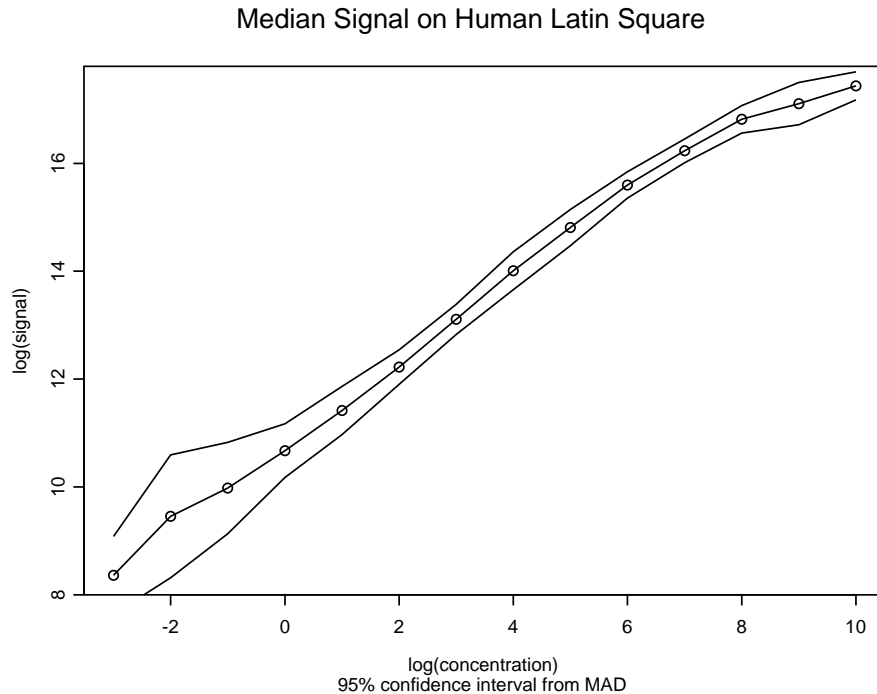


FIGURE 1. Logarithmic transformation approximately stabilizes the variance of signal. The signal values for twelve transcripts across 41 experiments were used to obtain a median signal and a median absolute deviation (MAD) for the fitted errors. The outer lines are at $3 \cdot \text{MAD}$ in each concentration category. Note the near-linear behavior of the signal in concentration.

transcripts did not amplify correctly as determined by gel analysis, and one array had to be discarded due to experimental failure. Thus, the human data represents only 12 transcripts and 41 arrays. The array chosen for the assay was the Hu95A design, with 16 probe pairs per transcript. To balance the analysis, the discarded array data was imputed from the first replicate of the set of experiments.

The yeast test data was obtained from a specially designed array that contained all probes matching the specified transcripts. In particular, the sixteen probe pairs corresponding to the yeast genome array were used for the analyses presented here. Fourteen groups of eight transcripts each were chosen for hybridization, so that each experiment had eight transcripts at any given concentration. The same 14×14 cyclic latin square was used, with each group of targets cycling through all the concentrations. Again, several targets failed to amplify, and so data for only 97 transcripts was used in this analysis. Four replicates were done of the 14 experiments comprising a latin square, across three manufacturing lots of arrays, resulting in a total of 56 arrays. Transcript information may be found in Liu et al. [2002].

The data was background adjusted by subtracting the mean of the lowest 2% of the probes, and scaled to the 2%-trimmed mean of the average difference score. Note that algorithms based on $I_{ij,PM} - I_{ij,MM}$ values are left unchanged by subtracting a background

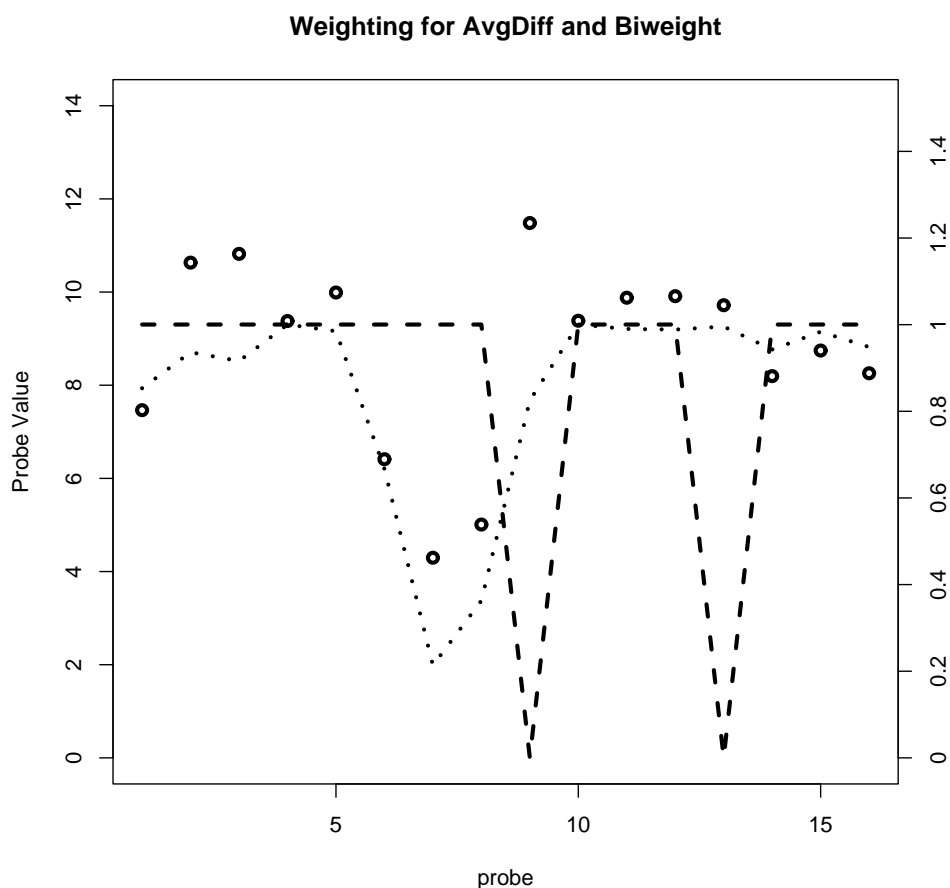


FIGURE 2. Probe values for a 16 pM spike distributed around the median value of 9.4 ($\text{AvgDiff} = 9.32$, $\text{Biweight} = 9.3$), showing the weights assigned by AvgDiff and Biweight to each probe pair. Note that probe pair 13 is an imputed value ($PM - MM = -847$) and was weighted by AvgDiff at zero, despite the PM intensity located in the middle of all PM values. Biweight assigns continuous weights ranging from 0 to 1 that depend on the distance from the median probe value, while AvgDiff assigns weights that are either 0 or 1.

value from both the perfect match and mismatch probe, so this does not compromise the power of such algorithms.

4. ON MISMATCHES

While not the central focus of the analysis, we tested to see if mismatches could be dropped from the biweight analysis, and could be replaced with a simple global estimate of background. Naturally, we cannot check all possible global estimates of background,

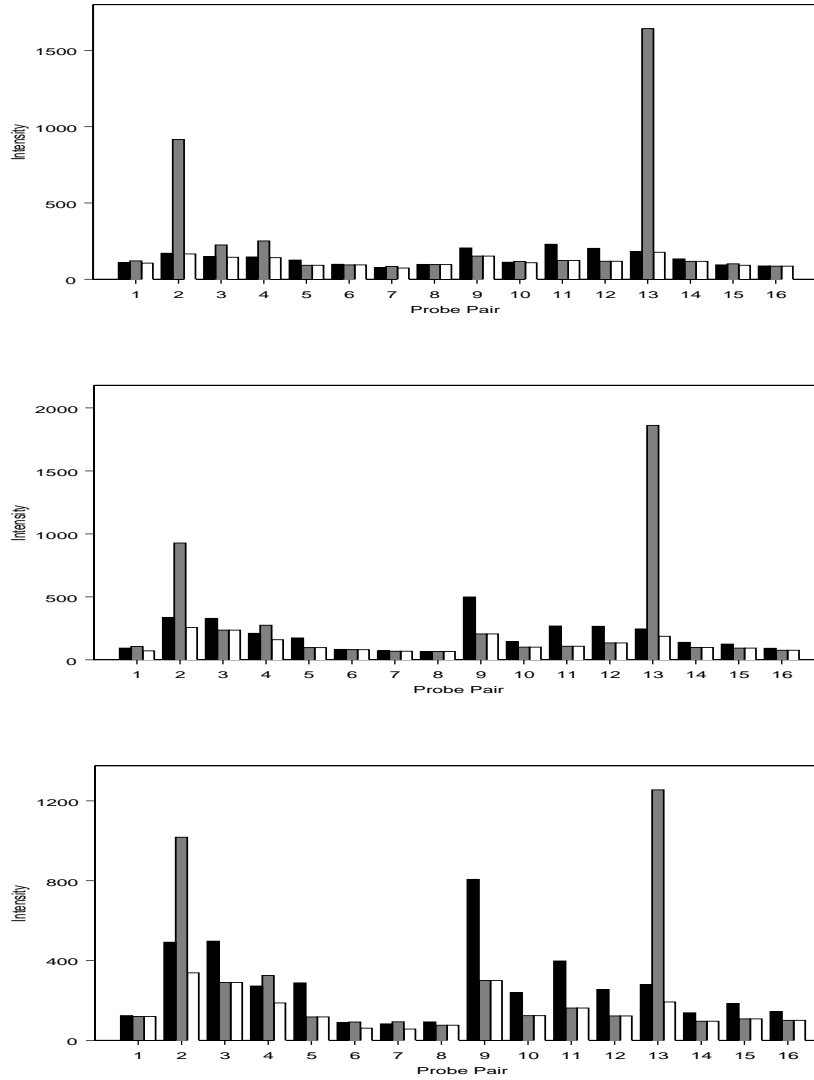


FIGURE 3. Perfect match (black), mismatched probes (grey), and estimates of stray signal (white) for three different concentrations, 0, 1, and 2 pM. Pictured are probes to transcript 36311_at from the Hu95A chip.

and we would be pleased to drop mismatched probes. We compared the performance using the above described 2% background against the performance with added mismatches. As seen in table 1, adding mismatches improves the correlation in the 0-2 pM range, and roughly halves the proportion of errors that occur across all transcripts. The performance of 16 probes, when split into 8 perfect match and 8 mismatch probes was marginally better than 16 perfect match probes. The performance with more probe pairs is, as expected, better at all concentration ranges. However, note that in the 2-32 pM range the correlation is

Kendall correlation (average)			
Probes	Concentration		
	0-2pm	2-32pm	32-512pm
8 PM	0.804	0.966	0.896
8 PM+8MM	0.884	0.9691	0.894
16 PM	0.863	0.979	0.942
16 PM+16MM	0.941	0.980	0.932

TABLE 1. MM probes improve performance at lowest concentration relative to simple background subtraction, but can degrade performance at higher concentrations.

only marginally improved by mismatches, and the performance of 16 perfect match probes is very similar to 16 probes split 8 and 8. Finally, in the 32-1024 range, mismatches are slightly detrimental and 16 perfect match probes has half the error rate of the 8 perfect match and 8 mismatch combination. Since we are most concerned with performance at the low concentration end, and the data sets under-represent the possible variation in stray signal across samples, we have chosen for our estimator the tradeoffs involved in continuing the use of mismatches.

REFERENCES

W.M. Liu, R. Mei, X. Di, T.B. Ryder, E. Hubbell, S. Dee, T.A. Webster, C.A. Harrington, M.-h. Ho, J. Baid, and S. Smeeckens. Analysis of high-density expression microarrays with signed-rank call algorithms. *Bioinformatics, this issue*, 2002.

AFFYMETRIX, 3450 CENTRAL EXPRESSWAY, SANTA CLARA, CA 95051
E-mail address: Earl.Hubbell@affymetrix.com