Majority Logic Decoding for Probe-level Microarray Data

Humberto Ortiz-Zuazaga[*], Tim Tully[†], and Oscar Moreno[‡]
[*]High Performance Computing facility, University of Puerto Rico, San Juan, PR, USA
[†]Dart Neuroscience LLC, San Diego, CA, USA
[‡]Gauss Research Laboratory, Inc., San Juan, Puerto Rico, USA

Abstract

Microarrays permit measuring thousands of genes in a single experiment. The analysis of these experiments is complex, because of multiple sources of experimental error in the system. We had previously described a method of coarse discretization and error correction for microarray data based on majority logic decoding. We extend this method to work with "probe level" microarray data, where each gene is represented by multiple probes. We have applied the error correction procedure to two data sets, one Affymetrix, one NimbleGen. The experiment is designed to validate analysis procedures by examining the degree of concordance the procedures produce across the data sets.

Our principal contribution is a technique to measure the concordance quantitatively. We have developed a technique based on mutual information to compare results obtained across the two data sets. This technique measures the concordance of the microarray expression measures across two different experiments, validating the results by confirming the same expression patterns occur in separate data sets collected using different techniques.

Our results show that our techniques result in a greater amount of shared information between data sets than traditional approaches based on averaging of probes and gene expression levels across repetitions. Coarse discretization yields a large improvement in concordance, and the error correction doubles the improvement again. Thus, these techniques have been proved superior to the traditional techniques in finding concordance between data sets, and thus, finding real changes in gene expression.

Keywords: microarray, error-correction, discrete methods, probe level data

1 Introduction

We have previously described a method for error correction of microarray data [1]. That method produces a coarse characterization of gene expression levels, based on majority logic decoding of thresholded genes from multiple repetitions. Many microarray experiments use multiple probes per gene. This is typical of Affymetrix style gene chips, but is also seen on oligo arrays. In the analysis of such data, a probe summarization step is performed. There exists a great variety of probe summarization techniques, many are compared in [2, 3].

Section 2.2 describes an extension of our discretization and error correction procedure to deal with multiple probes per gene. Section 2.3 describes our principal contribution, a technique based on mutual information to compare the degree of concordance of results obtained from two data sets. Mutual information had been used previously to perform reverse engineering of gene expression networks from microarray data [4] or cluster microarray data [5], but not to measure correlation across two data sets. Section 2.1 describes two data sets, obtained on two different microarray technologies that we use to validate our analysis procedures.

We have applied the discretization and error correction procedure to two *Drosophila* data sets described in Section 2.1, one Affymetrix and one NimbleGen, having multiple probes per gene. The experimental technique is designed to compare analysis procedures by measuring the degree of concordance in the two separate data sets. Analysis procedures that produce good concordance across the data sets are thus validated empirically. Our results show that our new technique results in a greater amount of shared information between data sets than traditional approaches based on averaging of probes and gene expression levels across repetitions (Section 3). We find much more correlation between the data sets than that detected by earlier techniques. This increased concordance is principally due to the recovery of many false negatives eliminated by the prior analysis techniques.

2 Methods

2.1 Microarray data

The data sets were produced in experiments comparing gene expression levels at different times after odor avoidance training of *Drosophila melanogaster* [6, 7]. The experiments were run on drosgenome1 chips from Affymetrix (Santa Clara, CA, USA), and a set of custom arrays from NimbleGen (Madison, WI, USA). A data set consists of 10 repetitions of each condition (massed training, spaced training) at 3 separate time points, 0 (no training), 6, and 24 hours after training. The Affymetrix arrays have 14 probes for each gene, and 14010 probe sets, including controls. The NimbleGen arrays have a set of probes with around 10 probes for each probe set, and 12240 probe sets, including controls.

These experiments were designed to test the degree of concordance between genes produced by different analysis techniques. Analyses that produce the same results across the two data sets should be detecting some ground truth of the biological system, and are less likely to be detecting spurious signal from the particular experimental technique.

2.2 Error correction of probe-level data

Our previous paper described an error correction method for replicate microarray data sets [1]. In that paper, we assumed each gene was represented by a single probe, as is typical of cDNA arrays. To extend our method to multiple probes, we first run our prior method on the data, treating each probe as a separate entity. This summarizes the repetitions, resulting in a set of calls for each probe. A call is "+" if a probe is upregulated compared to the control, "−" if the probe is downregulated compared to the control, "0" if the probe is within epsilon of the control, and "?" if the results are ambiguous. We then perform majority logic decoding on the set of probes corresponding to each gene as described in [1]. Briefly, in the Affymetrix data set, each gene is represented by 14 probes. If a set of probes has more than 7 symbols in agreement, we use that consensus symbol, otherwise we use "?" to denote an ambiguous call. Figure 1 illustrates the majority logic decoding results for an example probe set in the Affymetrix data. There are 14 rows and 4 columns, and in the result each column is set to the symbol occurring more than 7 times in the data.

```
[['0', '-', '0', '+'],
 ['0', '-', '+', '+'],
 ['0', '-', '0', '+'],
 ['0', '-', '+', '+'],
 ['0', '-', '+', '+'],
 ['-', '-', '0', '0'],
 ['0', '-', '0', '+'],
 ['0', '-', '0', '+'],
 ['-', '-', '-', '0'],
 ['0', '-', '0', '+'],
 ['-', '-', '0', '0'],
 ['0', '-', '0', '+'],
 ['0', '-', '0', '+'],
 ['0', '-', '+', '+']]

 'mld': ['0', '-', '0', '+']
```

Figure 1: Majority logic decoding of an example probe set

2.3 Sorting genes by weighted mutual information

Because we have two experiments performed on different microarray technologies, we wish to discover which genes demonstrate the same patterns of expressions in the two data sets. We developed a program to compare the expression in the two data sets using weighted mutual information. Mutual information can measure positive (same expression patterns) and negative (inverted expression patterns) correlation, but in our case, we want to select genes that show only very similar patterns of expression, not opposite patterns. Thus, we use a weighted variant of mutual information [8] as in Equation 1, where $w(x, y)$ is the weight assigned to the combination of symbol x and y, described below, $p(x, y)$ is the probability of the combination of symbol x and y, and $p(x)$ is the frequency of symbol x in the sequence X.

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} w(x,y) p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{1}$$

We set the weights such that similar patterns of expression are given higher weights, and opposite expression is given lower weight as shown in Equation 2.

$$w(x,y) = \begin{cases} 1.0 & \text{if } x = y,\ x, y \neq ? \\ 0.5 & \text{if } x = ? \text{ or } y = ? \\ 0.1 & \text{otherwise} \end{cases} \tag{2}$$

Ambiguous calls are given an intermediate weight.

Figure 2 illustrates the calls in each of the data sets for the example probe set. Equation 3 shows the computation of the weighted mutual information (WMI) for these two sequences.

```
ac = ['0', '-', '0', '+']
nc = ['0', '-', '-', '-']
```

Figure 2: Example calls for a single probe set in both data sets.

$$
\begin{aligned}
I(\texttt{ac},\texttt{nc}) = & \\
& w(\texttt{-},\texttt{-})p(\texttt{-},\texttt{-}) \log\left(\frac{p(\texttt{-},\texttt{-})}{p(\texttt{-})p(\texttt{-})}\right) + \\
& w(\texttt{0},\texttt{-})p(\texttt{0},\texttt{-}) \log\left(\frac{p(\texttt{0},\texttt{-})}{p(\texttt{0})p(\texttt{-})}\right) + \\
& w(\texttt{+},\texttt{-})p(\texttt{+},\texttt{-}) \log\left(\frac{p(\texttt{+},\texttt{-})}{p(\texttt{+})p(\texttt{-})}\right) + \\
& w(\texttt{0},\texttt{0})p(\texttt{0},\texttt{0}) \log\left(\frac{p(\texttt{0},\texttt{0})}{p(\texttt{0})p(\texttt{0})}\right) \\
= \quad & 1 \cdot 1/4 \cdot \log\left(\frac{1/4}{1/4 \cdot 3/4}\right) + \\
& 0.1 \cdot 1/4 \cdot \log\left(\frac{1/4}{2/4 \cdot 3/4}\right) + \\
& 0.1 \cdot 1/4 \cdot \log\left(\frac{1/4}{1/4 \cdot 3/4}\right) + \\
& 1 \cdot 1/4 \cdot \log\left(\frac{1/4}{2/4 \cdot 1/4}\right) \\
= \quad & 0.35 \qquad (3)
\end{aligned}
$$

We obtained from Affymetrix a file with the sequence annotations for every probe on the drosgenome1 chips, `DrosGenome1.na21.annot.csv`. We used the "Probe Set ID" and "Ensembl" columns to construct a map from the ID used by Affymetrix to the IDs used in the NimbleGen arrays. Several Affymetrix Probe Set ID have more than one Ensembl ID listed. Because of this we average the WMI for all NimbleGen probe sets that map to the same Affymetrix probe set.

We sum the average WMI over all Affymetrix probe sets, and obtain a single score for a particular analysis method, the summed weighted averaged mutual information or SWAMI.

With the SWAMI score, we can perform many analyses, and compare the SWAMI score obtained to determine which analysis technique produces the best agreement between the two data sets.

In addition, we sort the probe set list by the weighted averaged mutual information, this produces a list of probe sets ranked according to how informative and how similar they are between data sets.

2.4 Normalization and summarization tests

We set up a series of analyses to test the effect of different transformations and summarization algorithms on the concordance between the two data sets, as measured by the SWAMI score. We use the `affyPLM` package from BioConductor [9]. We set up a comparison of "log2", "sqrt" and "cuberoot" transformations on the expression values, and "Huber" "fair" and "Cauchy" methods of robust regression of the probe values. Once we produce the summarized data, we use the `limma` package from BioConductor to produce a discretization using the `decideTests` function [10, 11]. These discretizations are compared between the two data sets using the SWAMI score, just as we compared the error correction methods above. We also ran our error correction and clustering procedures on the data summarized using the `rma` command from BioConductor [12].

3 Results

Table 1 summarizes the total SWAMI scores obtained for several different transformation and regression methods on the *Drosophila* data. The defaults for `affyPLM` are log2 transformation and Huber regression, but sqrt transformation and fair regression yielded much better SWAMI scores on our data.

Table 1: SWAMI scores for several transformation and regression methods.

Transformation	Regression	SWAMI
log2	Huber	182
log2	fair	186
log2	Cauchy	169
sqrt	Huber	212
sqrt	fair	230
sqrt	Cauchy	200
cuberoot	Huber	207
cuberoot	fair	216
cuberoot	Cauchy	202

Table 2 presents the SWAMI scores for our error correction techniques on the *Drosophila* data.

All these scores are more than an order of magnitude higher than the scores for the `affyPLM` based methods. The highest score is the "trimmed mean" method, which discards repetitions which deviate most from the mean. In our case we discard 2 and keep 8 repetitions for each probe.

Table 2: SWAMI scores for error correction methods.

Method	SWAMI
trimmed mean	3657
mean	2535
consensus	3058
consensus vs mean control	1525

For comparison, Table 3 presents the SWAMI scores for our prior error correction scheme, using standard RMA to summarize probes [12]. The additional level of error correction afforded by the probes results in an increase of the SWAMI score.

Figures 3 and 4 show the frequency of the individual WMI scores per probe set for two representative methods, the sqrt-fair method, which obtained the best SWAMI score in Table 1, and the trimmed mean, the best performer in Table 2. The error correction methods show a wider distribution, whereas the `affyPLM` methods are very narrowly distributed around 0.

Figures 5 and 6 illustrate the distribution of nonzero calls in each probe set in the data. In binary vectors this would be called the "weight" of the vector. The sqrt-fair method produces calls distributed normally with mean around 5, whereas the trimmed mean method produces calls with nearly all entries nonzero.

4 Discussion

We have presented a novel method of error correction for probe-level microarray data, such as that generated by Affymetrix chips. We have also developed a novel scoring method for measuring the degree of agreement between two independent data sets representing the same or similar genes.

Table 3: SWAMI scores for RMA summarized data for several methods.

Method	SWAMI
trimmed mean	2753
mean	2610
consensus	1920
consensus vs mean control	1888

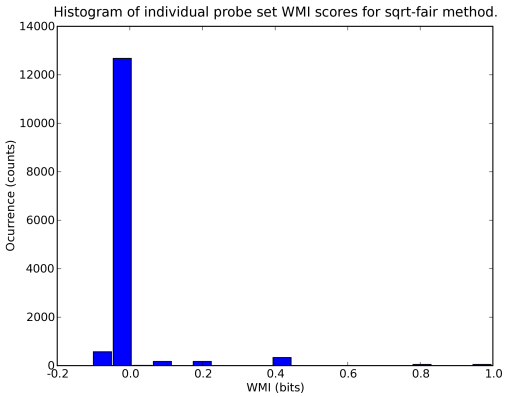

Figure 3: Distribution of weighted mutual information scores for individual probe sets in the sqrt-fair method.

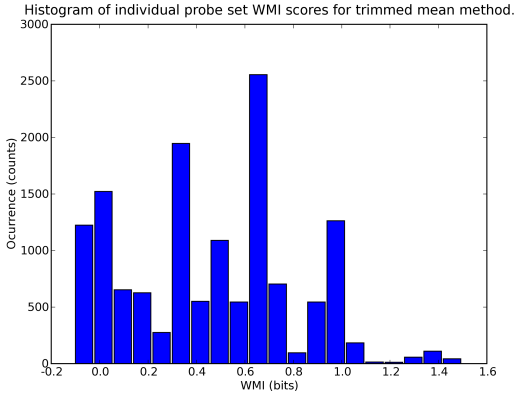

Figure 4: Distribution of WMI scores for individual probe sets in the trimmed mean method.

The SWAMI score measures mutual information between data sets, but is weighted by a score to produce biologically meaningful correlation, two sequences cannot be inversely correlated and still have a high SWAMI score.

We have applied our methodology to a large data set obtained from an odor-avoidance training experiment with *Drosophila melanogaster*. This experiment was designed to test different data analysis techniques by measuring concordance between the results on both data sets. The results indicate that our error correction procedure results in much higher SWAMI scores between two different data sets than other more common analysis techniques (Tables 1 and 2). The first reason is that the set of calls produced by the other techniques have more "0" calls, no significant change in expression, as seen in Figure 5 where the weight peaks

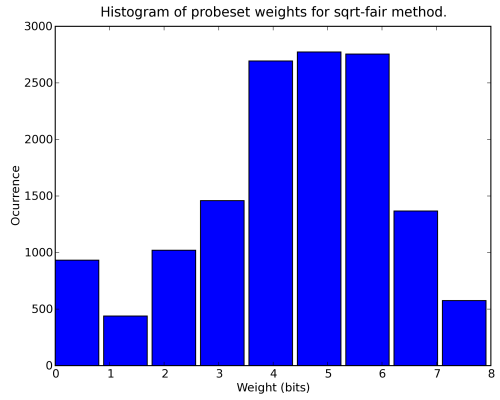

Figure 5: Distribution of nonzero calls for individual probe sets in the sqrt-fair method.

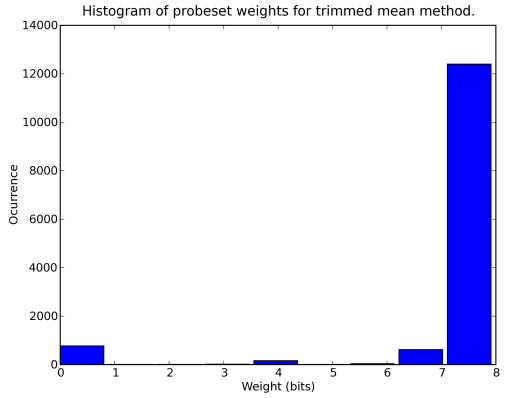

Figure 6: Distribution of nonzero calls for individual probe sets in the trimmed mean method.

around 5. The majority logic decoding results in calls with more "+" and "−" values, the weight illustrated in Figure 6 is much higher. If these calls were not in agreement between data sets, however, the SWAMI scores would not be high. Thus the second improvement is that the error correction procedure increases the degree of concordance between the data sets, as measured by the SWAMI score. Prior studies of concordance across microarray technologies have demonstrated poor results, the authors concluded diverse array technologies cannot be compared [13]. However, newer studies have shown large variability between labs, and that the best labs have low variability, even using different technologies[14].

The SWAMI score leads to a list of probe sets that in some sense optimize two biologically relevant criteria: the probe sets must be informative in the sense that they take on a range of values, and the probe sets must be consistent between the two data sets. These are precisely the kind of genes we seek to understand the molecular changes underlying the conditions we are studying.

The error correction and SWAMI procedures produce a list of probe sets sorted by the WMI score, there are a small number of probe sets with maximal scores. Future work should include examining these candidate probe sets to confirm their role in learning and memory processes.

5 Acknowledgments

The authors received partial support from a SCORE grant (S06GM08102) and an INBRE grant (P20RR16470) from the National Institutes of Health.

References

[1] H. Ortiz-Zuazaga, S. Peña de Ortiz, and O. Moreno de Ayala, "Error correction and clustering gene expression data using majority logic decoding," in *Proceedings of The 2007 International Conference on Bioinformatics and Computational Biology (BIOCOMP'07)*, Las Vegas, Nevada, USA, June 25-28 2007.

[2] L. M. Cope, R. A. Irizarry, H. A. Jaffee, Z. Wu, and T. P. Speed, "A benchmark for affymetrix genechip expression measures." *Bioinformatics*, vol. 20, no. 3, pp. 323–331, 2004.

[3] R. A. Irizarry, Z. Wu, and H. A. Jaffee, "Comparison of affymetrix genechip expression measures." *Bioinformatics*, vol. 22, no. 7, pp. 789–794, 2006.

[4] S. Liang, S. Fuhrman, and R. Somogyi, "Reveal, a general reverse engineering algorithm for inference of genetic network architectures." *Pac Symp Biocomput*, pp. 18–29, 1998.

[5] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements." *Pac Symp Biocomput*, pp. 418–429, 2000.

[6] J. C. Yin, J. S. Wallach, M. Del Vecchio, E. L. Wilder, H. Zhou, W. G. Quinn, and T. Tully, "Induction of a dominant negative CREB transgene specifically blocks long-term memory in *Drosophila*." *Cell*, vol. 79, no. 1, pp. 49–58, 1994.

[7] T. Tully, T. Preat, S. C. Boynton, and M. Del Vecchio, "Genetic dissection of consolidated memory in *Drosophila*." *Cell*, vol. 79, no. 1, pp. 35–47, 1994.

[8] S. Guiasu, *Information Theory with Applications*. New York: McGraw-Hill, 1977.

[9] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang, "Bioconductor: Open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, p. R80, 2004. [Online]. Available: http://genomebiology.com/2004/5/10/R80

[10] G. K. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments." *Stat Appl Genet Mol Biol*, vol. 3, p. Article3, 2004.

[11] ——, "Limma: linear models for microarray data." in *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, Eds. New York: Springer, 2005, pp. 397–420.

[12] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, "Summaries of affymetrix genechip probe level data." *Nucleic Acids Res*, vol. 31, no. 4, p. e15, 2003.

[13] W. P. Kuo, T.-K. Jenssen, A. J. Butte, L. Ohno-Machado, and I. S. Kohane, "Analysis of matched mRNA measurements from two different microarray technologies." *Bioinformatics*, vol. 18, no. 3, pp. 405–412, 2002.

[14] R. A. Irizarry, D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. N. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S. Q. Ye, and W. Yu, "Multiple-laboratory comparison of microarray platforms." *Nat Methods*, vol. 2, no. 5, pp. 345–350, 2005.