A Distribution Free Summarization Method for Affymetrix GeneChip® Arrays

Zhongxue Chen[1,2], Monnie McGee[1,*], Qingzhong Liu [3] , and Richard Scheuermann [2]
[1]Department of Statistical Science, Southern Methodist University, Dallas, TX 75275 USA, [2]Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX 75390 USA and [3] Department of Computer Science, New Mexico Institute of Mining and Technology, Socorro NM 87801 USA

ABSTRACT

Motivation: Affymetrix GeneChip brand arrays require a summarization step in order to combine the information in a probe set into one value representing the expression level of the corresponding gene. Here we present a new summarization method, Distribution Free Weighted (DFW) fold change, that uses the information of fold change but does not make any distributional assumptions for the data.

Results: Based on spikein data sets, we compare DFW with several popular methods, via both our own calculations and the 'Affycomp II' competition. The results show that DFW outperforms other methods when sensitivity and specificity are considered simultaneously. In fact, the area under the Receiver Operating Characteristic (ROC) curve for DFW is nearly 1.0 (a perfect value). Furthermore, DFW can obtain all the true positives with a small number of false positives. It is also computationally faster than most methods in current use.

Availability: The R package for DFW is available upon request.

Contact: mmcgee@mail.smu.edu

1 INTRODUCTION

[*]To whom correspondence should be addressed.

With the help of microarrays, researchers can measure the expression levels for tens of thousands genes simultaneously. This provides an opportunity for scientists to investigate the relationship between the functions of biological organisms and their genes at a genome-wide level. Of the several types of microarrays, the Affymetrix GeneChip® is the most widely used.

An Affymetrix GeneChip® can contain from six thousand to more than fifty thousand probe sets (genes), depending on the organism and platform. Each gene is interrogated by a set of probe pairs. Usually the number of probe pairs within a probe set is between 11 and 20. For each probe pair, there are two probes. A perfect match (PM) probe is a segment of a gene with a length of 25 nucleotides, which is perfectly complementary to a subsequence for the target mRNA. A mismatch (MM) probe is identical to the corresponding PM probe except that the middle (13th) base is intentionally changed to its Watson-Crick complement. MM probes were originally designed to measure the background of the corresponding PM probes.

The raw microarray data are usually highly “noisy”. Consequently, before any high level analysis, such as gene selection, classification, or clustering, is executed, a series of preprocessing procedures must be performed. These preprocessing steps can profoundly affect the results of high-level analyses. A typical preprocessing procedure consists of three steps: background correction, normalization and summarization, not necessarily in this order. The background correction step is typically done in an attempt to remove nonspecific binding; the normalization step reduces systematic variation between chips and the summarization step generates an expression value for each gene.

In this paper, we focus on the summarization step. There are several summarization methods in common use. The earliest one is the Affymetrix Microarray Suite (MAS 4.0) and later replaced by MAS 5.0 (Affymetrix, 2002). MAS 4.0 takes the average of the background corrected intensities of PMs within a probe set by removing the smallest and largest values (AvDiff). MAS 5.0 uses 1-step Tukey Biweight method to get a gene expression summary. Model Based Expression Index (MBEI, Li and Wong, 2001a, b) uses a model to estimate the signal based on the original scale. Robust Multi-chip Average (RMA, Irizarry et at., 2003a, b; Bolstad et al., 2003), uses median polish to obtain a single gene expression value for each probe set based on the logarithm-transformed intensities.

A recently developed summarization procedure, Factor Analysis for Robust Microarray Summarization (FARMS, Hochreiter et al., 2006), is also a model-based method that uses logarithm-transformed data. Model based methods are heavily dependent on model assumptions, and require estimation of model parameters in order to work. In practice, these assumptions may not be always appropriate for microarray data. Furthermore, parameter estimation is not an easy task for microarray data. Maximum likelihood procedures are typically unstable, and EM-based algorithms too slow, due to the large amount of data generated by a typical microarray experiment (Bolstad, 2004).

We propose a new nonparametric summarization technique, Distribution Free Weighted (DFW) fold change based method. In its current implementation, no background correction is performed, and quantile normalization, as in RMA and FARMS, is employed for normalization purposes. Furthermore, only the PM probes are used. We compare our new method with MAS 5.0 and its later improved version Probe Logarithmic Intensity Error (PLIER, Affymetrix, 2005), MBEI, RMA, RMA-noBG, Gene Chip RMA (GCRMA, Wu et al., 2004) and FARMS. We use two sets of Affymetrix Spikein data (available at: http://www.affymetrix.com) along with the "GoldenSpike" dataset (Choe et al., 2005), for comparison. There are two versions for FARMS: "l.farms" and "q.farms". Cyclic loess normalization is used for l.farms, while q.farms uses quantile normalization (Hochreiter et al., 2006). For the spikein data sets, "q.frams" performs better; therefore, we use "q.farms" only for comparison throughout this paper. The results show that DFW outperforms other methods when both of sensitivity and specificity are considered.

DFW and the method comparisons are implemented in R (Ihaka and Gentleman, 1996) and Bioconductor (Gentleman, et al., 2004). Both programs are available at http://www.bioconductor.org.

2 METHODS

It is known that approximately 30% of MM values are greater than PM values, and this has been found to be true for many Affymetrix platforms (Irizarry, et al., 2003b). In addition, the PM and MM values for the same transcript are highly correlated with each other, indicating the presence of non-specific hybridization. In other words, even if a target sequence is not perfectly complementary to a probe, it still can hybridize to that probe. Some small target sequences (for example, less than 13 nt), have the capability to hybridize to a PM and the corresponding MM. Non-biological variations can also be introduced during the steps of sample preparation. Furthermore, Li and Wong (2001a, b) found that the hybridization capabilities for PM probes within a probe set are not the same. They termed this difference in behavior the “probe effect”.

Since it is true that different probes within the same probe set hybridize with different strengths to the same target, a preprocessing method should take these differences into account. However, for most preprocessing methods, the probe effect for a probe is assumed to be a constant (Li and Wong, 2001a, b). It is well accepted that there is a linear relationship between the specific hybridization intensity and the concentration of the target mRNA (Lockhart et al., 1996). Under this assumption, the fold changes (ratios) of the specific hybridization intensities under different conditions from all PM probes within a probe set should be the same. Therefore, by considering the fold changes instead of the intensities, we can avoid the difficulty that comes from the probe effects issue. The spikein data sets are based on this linear relationship assumption (at least we should accept the assumption that there is a linear relationship for log intensity and log concentration) (Lockhart et al., 1996).

However, we can only estimate the fold changes between experiments based on the observed intensities that contain noise. There are several reasons why the estimated fold changes, even for probe-pairs that are part of the same probe set, are disparate from one another. First, no matter what methods we use, we cannot remove all the background noise. The background noise has a large effect on the estimated fold change, especially when the intensities are low. Second, the effect of nonspecific hybridization is different from probe to probe. Currently, we have no reasonable method to remove nonspecific binding from the observed intensities. Third, some PM probes are not really PM probes of the gene (probe set) assigned although they were thought to be so when the chip was designed (Harbig et al., 2005; Dai et al., 2005); this may be due to the lack of the knowledge for that gene at that time. Therefore, each PM probe should not be treated equally.

A good summarization method should not only utilize the information from multiple chips, as RMA, MBEI and FARMS do, but also consider the different qualities of PM probes within a probe set. The new method, DFW, is a multi-chip method and takes advantage of information among arrays. The "hybridization quality" of each PM probe within a probe set is estimated based on the fold change for that probe across all arrays. The final estimated fold changes are weighted averages by giving larger weights to high quality PM probes.

More specifically, the observed intensities are first logarithm-transformed to obtain the estimated relative (relative to 0) fold changes (on log scale) for each PM probe across arrays. The range (maximum – minimum) of relative fold changes for each PM probe is taken, and the median of the ranges of relative fold changes for PM probes within a probe set is calculated. We denote the median as M. The median-centered difference of the fold change range for PM probe i is denoted by x_i. We denote the maximum absolute value of x_i's as Max. The weighting function has the following form:

$$w(x) = (1 - (\frac{x}{Max})^2)^2 . \qquad (1)$$

And the final weight for probe i is:

$$w_i = \frac{w(x_i)}{\sum_j w(x_j)} . \qquad (2)$$

If all the ranges are the same, then each PM probe has the same weight. By using this weighting procedure, we usually give small or zero weights to those PM probes with poor qualities. Here we assess the quality of PM probes across all arrays, as this avoids a common situation where a PM probe may perform well for some arrays or conditions (for example, when the concentrations are high), but has poor behavior for other arrays or conditions.

The expression values (log base 2) of a probe set across arrays are from the weighted relative fold changes that are calculated based on the relative fold changes for each probe and its weight within that probe set. Usually, the differentially expressed genes have larger ranges of fold changes than that of non-differentially expressed genes. Therefore, the standard deviation (SD) of relative fold changes across arrays for a gene provides additional information. Differentially expressed genes would be expected to have larger

SD of fold changes than that of non-differentially expressed genes. The DFW summarization method uses information from both the range and SD of fold change.

First, the weighted relative fold changes (a vector with the length of the number of arrays) and the corresponding weighted range (WR, a scalar) are calculated based on the relative fold changes and the weight from each probe. The weighted relative fold changes are linearly transformed to be between 0 and 1 to give the transformed fold changes (TFC). A weighted standard deviation (WSD) is calculated in the same way as the weighted range. The final relative fold changes (expression values on log base 2 scale) for a gene across arrays are given by the following formula:

$$C_1 + C_2 \times TFC \times WR^m \times WSD^n \, . \qquad (3)$$

Here m and n are positive numbers, and the default values are set to be m=3, n=1 in DFW. The constant C_1 is the minimum value of the weighted relative fold changes before the linear transformation and constant C_2 is 0.01 by default. Neither C_1 nor C_2 affects the results of the comparisons in the next section.

3 RESULTS

3.1 Data sets

We compare our new method with others by using three publicly available spikein data sets (Dataset A, B and C). Dataset A is the Affymetrix Latin Square spike-in experiment done on the HG-U95Av2 array. For details on this experiment, see the

Affymetrix website (http://www.affymetrix.com). Our comparisons use 16 spikeins instead of the original 14, following the recommendations of Cope, et.al, 2004. Dataset B is the Affymetrix Latin Square spikein experiment performed on the HG-U133A array platform. It was originally designed with 42 spikeins (14 spikein groups with each group of three). McGee and Chen (2006) recently found that there are 22 additional spikeins in Dataset B. Therefore, there are actually 64 spikeins for Dataset B. We do our comparisons with both the original 42 spikeins, via the Affycomp II competition, and the 64 new spike-ins. Dataset C consists of six DrosGenome1 chips (two conditions with three replicates for each) with 3860 probe sets that can be detected as presented (Choe, et al., 2005). Among the 3860, 1309 have known fold changes from 1.2 to 4 and the remaining 2551 have the same concentrations under both conditions (spikein and control).

3.2 Results

The "Affycomp II" competition (http://affycomp.biostat.jhsph.edu; Cope et al, 2004; Irizarry et.al, 2006), allows comparisons among fifty-four and fifty-five (at the time this paper was prepared) public competition methods based on data sets A and B, respectively. The competition uses many comparison statistics, but only the various area under the ROC curve (AUC) statistics are not scale-dependent (Hochreiter et al., 2006). Furthermore, the AUC allows comparison based on sensitivity and specificity, which are the most important characteristics of a preprocessing method from the standpoint of a researcher. DFW outperforms all others with regard to sensitivity and specificity simultaneously. Weighted AUC values, as calculated on the Affycomp II website, for DFW, RMA, MAS 5.0, MBEI, FARMS and GCRMA, for both Datasets A and B, are given in Table 1. The weighted AUC values for DFW are 1 from both datasets A and B. Note that this comparison uses 42 spikeins for Dataset B. In Table 2, we show the results using all of the 64 spikeins for Dataset B (McGee and Chen, 2006). In addition, we take into account the fact that some experiments are designed to demonstrate much larger differences between concentrations of the spike-in genes than are others.

Table 1. Average AUC for datasets A and B different methods

Dataset	DFW	FARMS	GCRMA	RMA	RMA-noBG	MAS5	MBEI	PLIER
A	**1.00**	0.91	0.69	0.60	0.65	0.05	0.26	0.03
B	**1.00**	0.95	0.57	0.65	0.63	0.06	0.40	0.20

Table 2. AUC comparison for Dataset A (#FP=5)

d	DFW	FARMS	GCRMA	RMA	RMA-noBG	MAS 5	MBEI	PLIER
1	1.000	0.692	0.566	0.453	0.587	0.063	0.063	0.046
2	1.000	0.849	0.793	0.764	0.784	0.126	0.206	0.079
3	1.000	0.865	0.865	0.882	0.864	0.188	0.448	0.066
4	1.000	0.933	0.902	0.927	0.933	0.429	0.598	0.067
5	1.000	0.936	0.982	0.992	0.975	0.746	0.708	0.040
6	1.000	0.937	0.993	0.996	0.998	0.848	0.763	0.000
7	1.000	0.937	0.996	0.998	0.997	0.862	0.801	0.000

For the Latin Square data sets, we compared pairs of experiments that were separated by the same number of permutations (where d = number of permutations), of the Latin Square. See McGee and Chen (2006) for a more complete explanation of d. Usually, it is harder to detect the true positives for small d than for large d.

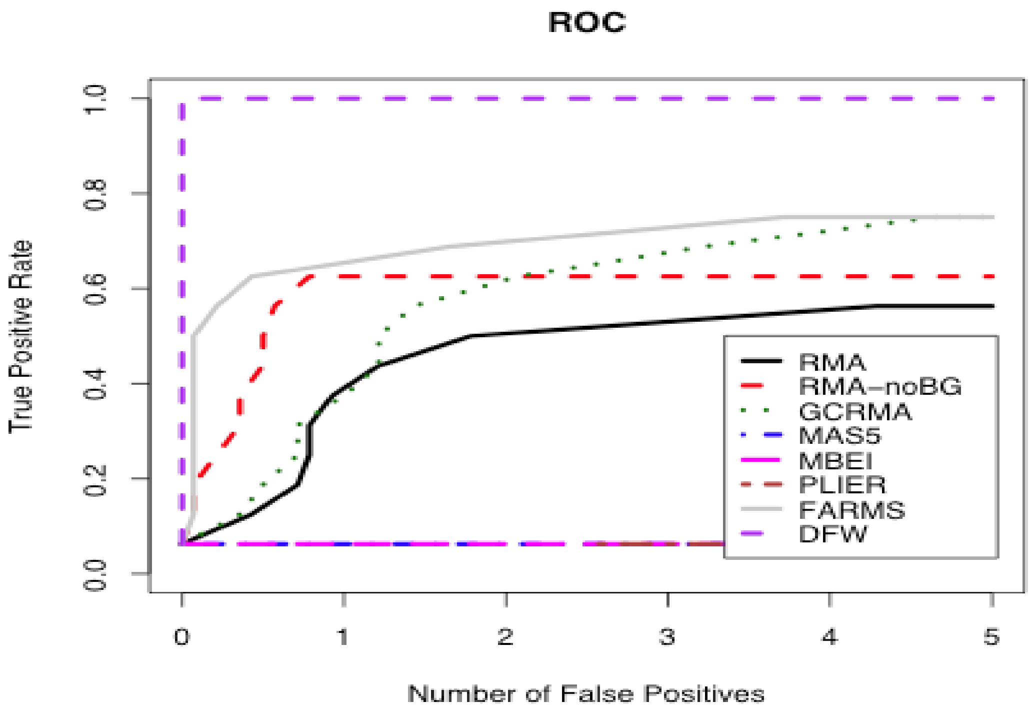

Figure 1. ROC plots based on Dataset A when d=1 and number of false positives is 5. DFW obtains all the true positives without any false positive.

The AUC was calculated for a cutoff of various numbers of false positives (e.g. the number of false positives is 5 for Table 2). The values are then standardized so that the area is between 0 and 1. For Dataset C no cutoffs were set since the fold changes for this dataset are usually very small and many spikeins can not be detected as differentially expressed genes for a small number of false positives.

Figure 1 is the ROC curve plot of Dataset A for d=1 when the number of false positives is 5. The DFW method can obtain all of the true positives without any false positives and the curve is above any other curves from other methods. For d = 2, 3, … , 7, we obtain the similar plots for ROC curves. In other words, in all situations, DFW outperforms the other methods.

Table 3. AUC from Dataset A for given numbers of false positives when d=1

# FP	DFW	FARMS	GCRMA	RMA	RMA-noBG	MAS 5	MBEI	PLIER
2	**1.000**	0.626	0.362	0.320	0.530	0.063	0.063	0.000
5	**1.000**	0.692	0.566	0.453	0.587	0.063	0.063	0.046
10	**1.000**	0.754	0.658	0.529	0.638	0.063	0.103	0.054
20	**1.000**	0.819	0.729	0.606	0.663	0.063	0.253	0.058
40	**1.000**	0.872	0.788	0.646	0.675	0.063	0.416	0.060

Table 2 gives the values for the AUC based on Dataset A when the number of false positives is 5. For all d, the AUC from DFW is 1 while some of the other methods have very small values, especially for small d. Table 3 gives AUC values for d=1 of all methods when the false positives are 2, 5, 10, 20 and 40. The AUC from DFW consistently obtains the best value of 1. Based on Table 2 and 3, it is clear that our new summarization method, DFW, outperforms all other methods for all d.

Figure 2 is the ROC curve plot of Dataset B for d=1 when the number of false positives is 10. Sixty-four spikeins are used, instead of the original 42 (McGee and Chen, 2006). From the plot, the DFW method can obtain all of the true positives with a few false positive (less than 2 on average) and the curve from this method is above any other curves of others. For d equals to 2, 3, … , 7, we get almost similar plots for ROC curves (not shown here). In other words, in all situations, our new method, DFW outperforms others.

Table 4 gives the comparison results based on Dataset B when the number of false positives is set to be 20. Again, our new method DFW, can obtain all the true positives with a small number of false positives. For d = 2,3,…,7, DFW gives similar results. Table

5 gives the average AUC when d=1 and the numbers of false positives are 5, 10, 20, 50, 100. DFW always obtains a value of AUC close to 1 while some of the other methods obtain very small AUC values.

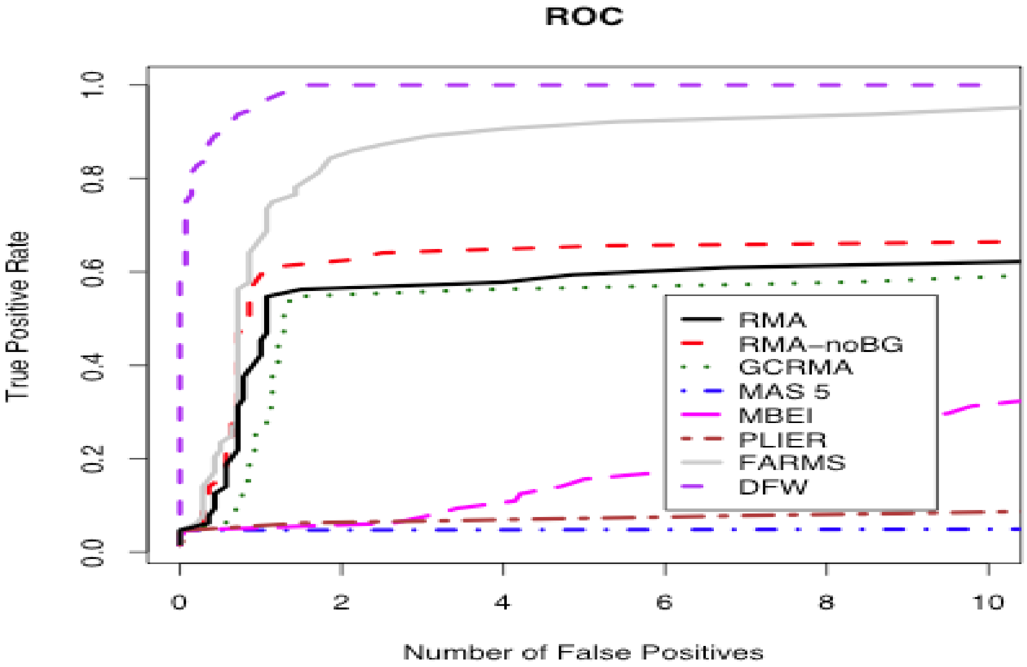

Figure 2. ROC plots based on Dataset B (with 64 spike-ins) when d=1 and number of false positives is 10. DFW obtains all the true positives with a few false positives.

Table 4 AUC comparison for Dataset B (#FP=10)

d	DFW	FARMS	GCRMA	RMA	RMA-noBG	MAS 5	MBEI	PLIER
1	**0.986**	0.841	0.514	0.550	0.607	0.047	0.155	0.070
2	**0.987**	0.864	0.589	0.674	0.635	0.106	0.363	0.197
3	**0.993**	0.932	0.658	0.806	0.815	0.276	0.484	0.434
4	**0.997**	0.947	0.783	0.884	0.915	0.514	0.649	0.658
5	**0.999**	0.970	0.934	0.948	0.961	0.760	0.762	0.822
6	**1.000**	0.978	0.955	0.965	0.969	0.9849	0.810	0.887
7	**1.000**	0.982	0.976	0.982	0.983	0.907	0.836	0.912

Table 5. AUC from Dataset B for given numbers of false positives when d=1

# FP	DFW	FARMS	GC-RMA	RMA	RMA-noBG	MAS 5	MBEI	PLIER
5	**0.971**	0.749	0.453	0.493	0.555	0.047	0.079	0.060
10	**0.986**	0.841	0.514	0.550	0.607	0.047	0.155	0.070
20	**0.993**	0.901	0.554	0.600	0.640	0.047	0.286	0.085
50	**0.997**	0.950	0.589	0.657	0.675	0.047	0.451	0.154
100	**0.999**	0.976	0.613	0.702	0.698	0.057	0.536	0.248

Table 6. AUC comparison for Dataset C

FC	Spike-ins	DFW	FARMS	GCRMA	RMA	RMA-noBG	MAS 5	MBEI	PLIER
1.2	172 (167)	**0.900**	0.878	0.860	0.863	**0.888**	0.534	0.814	0.115
1.5	182 (169)	**0.682**	0.659	**0.743**	0.486	0.595	0.307	0.406	0.100
1.7	181 (179)	0.623	**0.640**	**0.786**	0.449	0.567	0.153	0.459	.0321
2	146 (139)	**0.832**	0.783	**0.886**	0.746	0.799	0.213	0.761	0.524
2.5	192 (182)	**0.902**	0.879	**0.920**	0.848	0.880	0.306	0.872	0.619
3	97 (93)	**0.965**	0.948	**0.962**	0.928	0.947	0.423	0.946	0.734
3.5	184 (184)	**0.948**	0.941	**0.961**	0.940	0.944	0.596	0.942	0.850
4.0	177 (177)	**0.986**	0.975	0.985	0.978	0.982	0.633	**0.988**	0.878
all	1331	**0.846**	0.829	**0.882**	0.768	0.816	0.394	0.761	0.504

Table 6 is the comparison result based on GoldenSpike dataset (Dataset C). To calculate the AUC, we use the number of false positives necessary to obtain all of the true positives. The AUC values are then normalized so that the numbers are comparable among methods. The first column is the values of fold changes (spikein group vs. control group) and the second column gives the number of spikeins that have corresponding fold changes given in column 1. Some of the spikeins do not have the exact fold changes listed. For example, there are 167 out of 172 spikeins that have exact fold change of 1.2 but 5 of the 172 have fold changes greater than 1 but less than 1.2. Choe et al. (2005) explain this phenomenon in their paper. The values in the parenthesis of column 2 are the numbers of spikeins that have the exact fold changes indicated in column 1.

The first two largest values of AUC for each category of spikeins are highlighted in Table 6. We see that DFW and GCRMA almost always give the highest AUC values. Usually GCRMA has slightly larger values of the AUC than does DFW. This may be due to the fact that there is no background correction employed in the current implementation of DFW. Dataset C was created to mimic real data as closely as possible; therefore, it is noisier than the Latin Square data sets, and probably requires more background correction and normalization. Only minimal background correction and normalization is required for the Latin Square spikein data sets since they are designed to have little background noise.

Table 7: CPU time (in seconds) for various methods on the three spike-in data sets (the best times for each method are marked with *)

Method	Dataset A	Dataset B	Dataset C
RMA	361	400	152
RMA-noBG	365	366	148
GCRMA	240	221	76
MAS 5.0	1118	1116	131
PLIER	324	241	**17***
MBEI	847	272	272
FARMS	153	205	283
DFW	**121***	**150***	**69**

CPU time for each method, as given on a PowerMac G5 running R Cocoa GUI (Iacus and Urbanek, 2005) with R version 2.2.0, is given in Table 7. The computational time for DFW is much less than the other methods, such as RMA, MBEI, FARMS and PLIER. This is because DFW is not an iterative method. MBEI and FARMS, for example, require iterative algorithms to estimate the necessary model parameters. In the case of the GoldenSpike data, PLIER is faster than DFW. However, DFW is the second-fastest method, and it gives much more accurate results. In addition, it should be noted that MBEI does not always converge, even for the Latin Square data sets, and is particularly unsuitable for the GoldenSpike data set due to the small number of arrays.

4 CONCLUSION

We have proposed a new nonparametric summarization technique, distribution free weighted fold change based method (DFW). This method is compared with currently commonly used methods, based on the publicly available spikein data sets. Our new method outperforms others when sensitivity and specificity are considered simultaneously. In addition, DFW requires less computational time compared with others.

ACKNOWLEDGEMENTS

The authors wish to acknowledge support for this work from Dr. Milton Packer, Director, Department of Clinical Sciences, University of Texas Southwestern Medical Center. This research was supported by the National Institutes of Health contracts N01-AI40076 and N01-AI40041 to RHS.

REFERENCES

Affymetrix, Inc.. (2002) Statistical algorithms description document.

Affymetrix, Inc. (2005) Technical note: guide to probe logarithmic intensity error (PLIER) estimation.

Bolstad BM. (2004) *Low Level Analysis of High-density oligonucleotide array data: Background, normalization and summarization* [dissertation]. Department of Statistics, University of California at Berkeley.

Bolstad, B.M. et al. (2003) A comparison of normalization methods for high density oligunucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185-193.

Choe, S.E. et al. (2005) Preferred analysis methods for Affymetrix genechips revealed by a wholly defined control datasets. *Genome Biol*., 6, R16.1-R16.6.

Cope, L.M. et.al. (2003) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**, 323-331.

Dai, M. et al (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*., 33(20):e175; doi:10.1093/nar/gni179

Gentleman, R.C. et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.

Harbig, J. et al. (2005) A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. *Nucleic Acids Res*., 33(3):e31; doi:10.1093/nar/gni027.

Hochreiter, S. et al. (2006) A new summarization method for affymetrix probe level data. *Bioinformatics*, **22**, 943-949

Iacus, S.M. and Urbanek, S. (2005) R Cocoa GUI 1.14 (2129), S.M., © R Foundation for Statistical Computing.

Ihaka, R. and Gentleman, R.C. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat*,, **5**, 299-314.

Irizarry, R.A. et al. (2003a) Summarize of Affymetrix GeneChip probe level data. *Nucleic Acids Res*., **31**, 1-8

Irizarry,R.A. et al. (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249-264.

Irizarry, R.A. et al. (2006) Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, 22, 789-794.

Li, C and Wong, H.W. (2001a) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Nat. Acad. Sci*., **98**, 31-36.

Li, C and Wong, H.W. (2001b) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol*., **2**, research0032.1-0032.11.

Lockhart, D.J. et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol*., **14**, 1675-1680

McGee, M. and Chen, Z. (2006) New spiked-in probe sets for the Affymetrix HG-U133a Latin square experiment. *COBRA Preprint Series*, Article 5

Wu, Z. et.al. (2004) A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.*, **99**, 909-917.