Robust Estimation of Microarrays

Osorio Meirelles[1], Sushmita Roy[2] and Margaret Werner-Washburne[3]

Abstract — **The identification of differentially expressed genes using microarray data often assumes every microarray has equal value. This assumption can be inappropriate because of technical and biological variability among the arrays. To address this, we have developed a robust statistical unsupervised algorithm that estimates a weight for each array. The weight indicates the level of confidence in the array quality. Using a parametric iterative Empirical Bayes algorithm we estimated the array weights. Predictive gene scores were generated using the weights and then ranked, resulting in more accurate gene lists.**

Keywords **— Empirical Bayes, microarrays, robust, weight**

I. Introduction

Microarray data has many sources of variation that can affect the interpretation of the analysis. In some cases, gene expression outliers can lead to highly distorted results. In such cases the normal distribution assumption of gene expressions may be problematic. In order to address this, we have developed an algorithm in MATLAB, that is robust, easy to use and is also computationally efficient.

We propose a "robust estimator" weighting algorithm, which reduces the impact of outliers. The array weights are based on a similarity metric obtained by comparing each array with all the remaining arrays. In order to estimate the weights, the proposed method applies the original idea of Empirical Bayes estimation [1], and extends the weighting methodology for integrating similar datasets to microarray data [2,3]. Our approach is implemented via a parametric iterative algorithm called Empirical Bayes Robust Estimation of Microarrays (EBREM). Compared to existing approaches EBREM produces more accurate gene score predictions and generates more robust gene lists.

II. Methods

The algorithm was tested on a yeast dataset comprising 88 arrays measuring the gene expression difference under two experimental conditions [4]. After estimating the array weights, weighted and non-weighted gene scores were calculated using different types of scores and were later compared to one another. Next we selected gene lists of sizes 150, 200, 300, 400, 500, 600, 700, 800 for every type of score, based on top highest gene expression scores.

Two validation criteria were used, a statistical validation and a biological validation. The statistical validation used gene list overlap (defined as the proportion of the intersection) between a randomly generated array subset A of 4 arrays, and subset B, the remaining 84 arrays. The biological validation was evaluated by the p-value output generated from GO term finder [5].

III. Results

A total of 660 simulations were executed comparing the average gene overlap between subsets A and B.

After comparing gene overlaps for several gene score types, weighted and non-weighted, we find that all weighted versions of gene scores, had a higher gene list overlap than their non-weighted versions. This indicates that weighted scores have a higher stability than non-weighted scores.

gene list size	wgex	gex	wbin	bin	wt	t
150	0.706	0.673	0.533	0.489	0.168	0.141
200	0.698	0.669	0.573	0.535	0.207	0.174
300	0.693	0.663	0.610	0.581	0.283	0.238
400	0.691	0.658	0.631	0.604	0.345	0.292
500	0.691	0.659	0.655	0.629	0.388	0.337
600	0.694	0.659	0.662	0.647	0.426	0.381
700	0.697	0.667	0.667	0.646	0.460	0.415
800	0.705	0.671	0.666	0.636	0.493	0.449

Statistical validation of a small subset of tested gene scores. Gene scores shown are: **gex** (gene expression mean), **wgex** (weighted gene expression mean), **bin** (binary gene exp. transformation mean), **wbin** (weighted binary gene exp. mean), **t** (studentized gene exp.), **wt** (weighted studentized gene exp).

IV. Conclusion

We have presented a method, EBREM that exploits array weighting to robustly combine heterogeneous information across microarrays. All weighted score types outperformed their non-weighted versions. EBREM is also computationally efficient converging in approximately 1 second on a single processor laptop, Intel Pentium III, 1.13Ghz, whereas other weighting methods took several minutes to hours to complete.

References

[1] Herbert Robins (1956), Non Parametric Empirical Bayes (NPEB).
[2] Ibrahim and Cheng (2000), Power Prior Distributions for Regression Models.
[3] Richie and Smith (2006), Empirical quality weights in the analysis of microarray data.
[4] Aragon (2008), Characterization of Differentiated Quiescent and Nonquiescent Cells in Yeast Stationary-Phase Cultures.
[5] Elizabeth I. Boyle (2004), GO::TermFinder---open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes.

[1] Department of Mathematics and Statistics, University of New Mexico
[2] Department of Computer Science, University of New Mexico
[3] Department of Biology, University of New Mexico