Improved Identification of Differentially Expressed Genes Using Pareto Set based Pruning

Jianjun Hu[1], **Jia Xu**[1]
{jianjunh@cse.sc.edu, xj182904@gmail.com}
[1]Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, 29208, USA

Abstract - *Identification of differentially expressed genes (DEGs) from microarray datasets is one of the most important analyses for microarray data mining. Popular algorithms such as statistical t-test, fold change, and rank product can be improved by considering other features of differentially expressed genes. We proposed a parameter-free non-dominated Pareto set based gene pruning algorithm for pruning non-differentially expressed genes before applying standard DEG identification algorithms. All genes are mapped to a feature space composed of average differences of gene expression and average expression levels and it is observed that differentially expressed genes tend to be located in boundary regions. Experiments on 17 Gene Omnibus Database (GEO) datasets showed that Pareto gene pruning can significantly improve popular algorithms such as t-test, rank product, and fold change in terms of prediction accuracy and AUC values with improvements ranging from 11% to 50% in terms of the number of identified true DEGs.*

Keywords: differentially expressed genes, Pareto set, microarray, gene pruning, gene ranking.

1 Introduction

Microarray based identification of differentially expressed genes (DEG) are now routinely used by biologists. There are two main categories of DEG identification algorithms. The first type includes single gene testing approaches such as fold change [1], rank product [2], t-test and its variants [3]. These methods are characterized by a single statistics score used to rank genes from significantly differentially expressed genes to no-change ones. The second type includes gene set testing approaches such as gene set enrichment analysis [4;5]. These methods are featured by exploiting externally determined functionally related gene sets to test the significance of differential expression of a group of genes. Despite increasing usage of gene set analysis methods [6], single-gene based DEG identification algorithms still dominate the practice of biological differential gene expression analysis [7-10]. This is partially due to their simplicity and less requirement on gene annotation. Thus improving single-gene DEG identification algorithms still has a lot of implication for DEG microarray analysis.

A major issue of current DEG microarray analysis is the limited number of samples in most biological studies, which makes many statistical test methods ineffective [11;12]. This issue has been addressed recently using a few strategies such as gathering information across similar genes (Bayes t-test approach [13], local pooled error algorithm [14], and the famous SAM algorithm [15]) or using external information to improve variance estimation [16] [17].

Here we propose a Pareto set based gene pruning algorithm to improve DEG identification algorithms such as fold change, t-test, or rank product. Our pruning algorithm is based on patterns of true DEGs in the space composed of average difference of gene expression level between two classes and the average gene expression level. It is motivated by the observation that experimentally verified true DEGs for 38 real-world datasets tend to also have high expression levels [18]. The first step is to prune non-DEGs from the whole gene list based on the characteristics of experimentally verified differentially expressed genes. In the second step, statistics-based DEG identification algorithms such as t-test are applied to rank genes. The non-DEG pruning is able to enrich true DEGs in the remaining gene lists. This is especially desirable for small-size microarray datasets. It can also be used to greatly reduce the computational cost of DEG algorithms that search gene combinations [19] where all gene-pairs need to be ranked. Based on systematic evaluation on 17 real-world datasets with a total of 184 true DEGs applied to four existing DEG algorithms, we showed that the Pareto pruning can significantly improve the performance of these traditional algorithms. For example, the Pareto Pruning algorithm can prune 81% genes out of 22283 while keeping 87% true DEGs out of 184 for 17 datasets we tested. The enrichment of true DEGs in the pruned gene list is almost six times of the original list. Pareto pruning is also shown to improve the AUC score of rank product algorithm by up to 48% and helps it to find 47% (50) more true DEGs when the cutoff top K=550. For fold change and t-test, it can find 24% (43) and 10% (20) more true DEGs.

2 Methods

The main idea of Pareto pruning is that identification of DEGs can be improved by considering characteristics of experimentally verified DEGs. One such feature is that true

DEGs tend to have high expression values [20]. Specifically, differentially expressed genes are usually located in the boundary region in the 2-D feature space of average gene expression (AG) versus average difference of gene expression (AD). Fig.1. shows the distribution of true DEGs in the 2D space for four microarray datasets: GSE9499, GSE6342, GSE6740-1, and GSE6740-2 from GEO database [21]. It is found that true differentially express genes are located on the boundary regions with either high average expression difference or high expression levels. This motivates our idea to develop a Pareto gene pruning algorithm for pruning non-DEG genes.

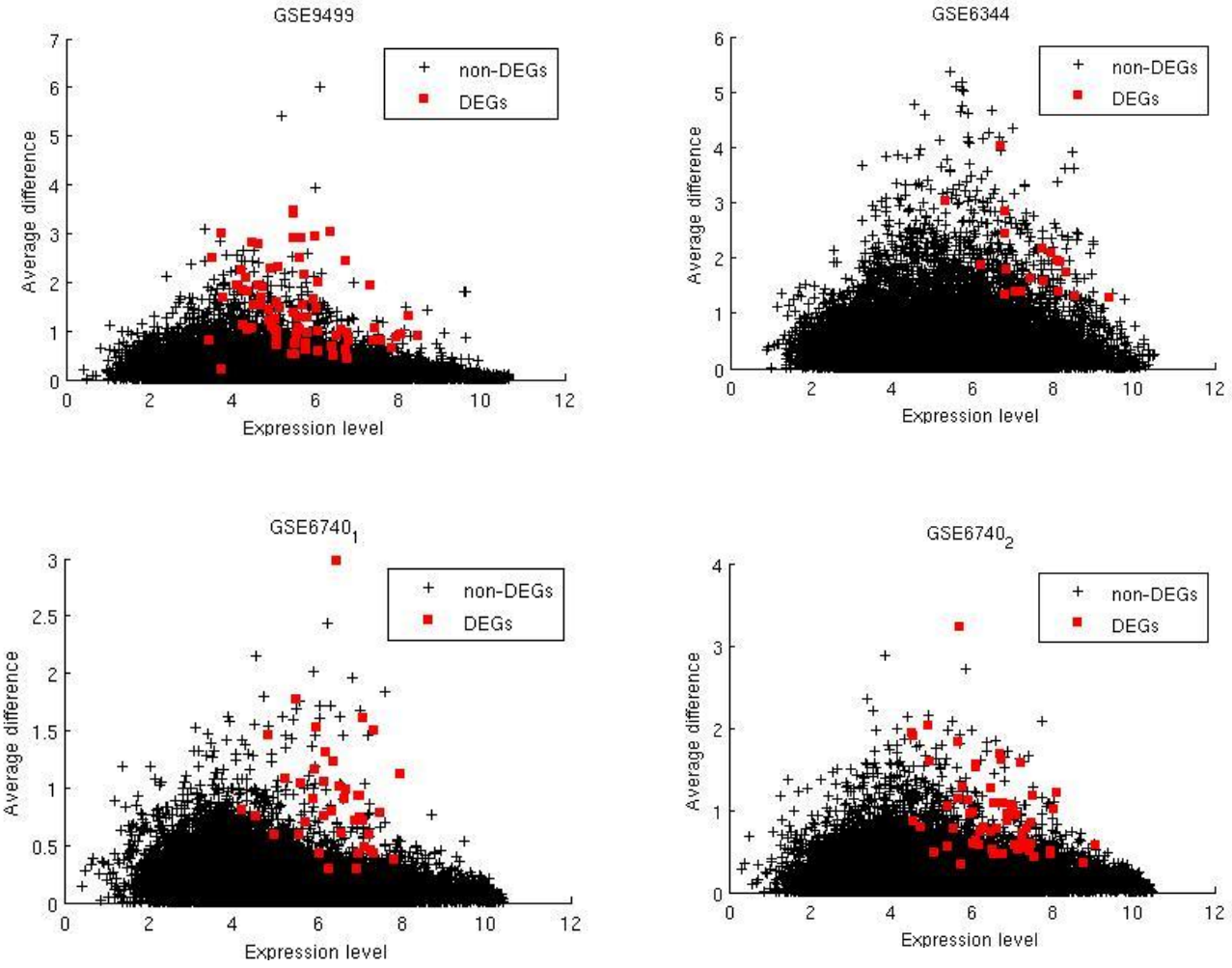


Fig.1 Distribution of verified differentially expressed genes in the expression-average expression difference (AG-AD) space for four datasets from GEO database. Differentially expressed genes tend to be located in the boundary areas.

2.1 Pareto gene pruning algorithm for DEG identification

The main idea of Pareto gene pruning is to remove non-DEGs that usually appear within the dense part of the AG-AD space (Fig 1). Assume M is a microarray matrix with N genes (rows) and P conditions (columns). P1 profiles of P correspond to condition A and P2=P-P1 profiles correspond to condition B. The average expression level (AG) of a gene X_i is defined as $(X_i^A + X_i^B)/2$, where X_i^A and X_i^B are the average expression level (log-scaled) of gene X_i under condition A and B. The average difference of gene expression of a gene X_i is defined as$|X_i^A - X_i^B|$. Since the expression values are log-transformed, the average difference here essentially represents the expression ratios.

It is observed that true DEGs tend to dominate non-DEGs in the AG-AD space. These non-dominated genes are called Pareto genes. A non-dominated gene is defined as one that there is no other genes which have greater values of both average gene expression level and average expression difference. The Pareto gene pruning algorithm works as follows. The input of Pareto gene pruning is the

number of genes K to keep. First, each gene X_i is mapped into the (AG, AD) feature space. Second, we compare each gene to remaining genes to check if there is any gene that dominates its (AG, AD) values, if not, it is added to the list of non-dominated genes. After first round of inspection of non-dominated genes, if the total non-dominated genes is less than K, these genes are removed from the pool and a next round of non-dominated gene screening is applied to generate second-level non-dominated genes until K non-dominated genes are collected. A major feature of this method is that there is no extra parameter to specify except the number of potential DEG candidates. After Pareto gene pruning, K non-dominated genes can then be ranked by conventional DEG algorithms such as the fold change, rank product, and t-test as describe below.

2.2 DEG Identification Algorithms

We applied three popular DEG identification algorithms to the 17 datasets with or without Pareto pruning.

- Fold Change (FC) is the most popular DEG algorithms among biologists. It ranks genes based on the ratio of average gene expression under two conditions. Usually a 2-fold change is regarded as significant in many biological studies. A major criticism of FC is that it doesn't consider the case that genes with low expression level in both conditions but with small variances can be ranked high. As shown by Figure 1, most true DEGs tend to also have high gene expression levels.
- Rank Product (RP) [22;23] ranks genes based on product of rank ratios for multiple A-B conditions. The results and simplicity of RP is similar to FC but overcomes its most significant limitations. It also provides a statistically rigorous estimation of significance. It was reported to have good performance for small datasets or noisy datasets.
- T-statistics (tTest) is one of earliest and still popular method used in DEG identification. The major advantage is that it can consider the variation of genes in its ranking. The limitation is that for small datasets, the estimation of gene expression variances is not reliable which can lead to bad performance.

3 Results

3.1 Data set preparation

We used 17 disease state or dose response analysis datasets of Homo sapiens out of the 36 GEO datasets collected by Kadota et al. [24] for comparing Weighted Average Difference (WAD) algorithm to other DEG algorithms. These datasets are provided with verified DEGs determined by real-time polymerase chain reaction (RT-PCR). These 17 datasets have been normalized and transformed into log scale. Out of the 17 datasets, only 7 have more than 10 samples for both conditions. Four datasets have less than 5 samples per condition. These datasets show that real-world GEO datasets, especially historical ones tend to have small sample sizes. The 17 Datasets cover a variety of biological or medical studies: GSE1462 (mitochondrial DNA mutations), GSE1615_1 (Valproic acid treatment), GSE1650 (chronic obstructive pulmonary disease), GSE2666_2(bone marrow Rho level effect), GSE3524 (tumor of epithelial tissue), GSE3860 (Hutchinson–Gilford progeria syndrome), GSE4917 (breast cancer), GSE5667_1 (atopic dermatitis), GSE6236 (Adult vs. fetal reticulocyte transcriptome comparison), GSE6344 (renal cell carcinoma disease), GSE6740_1 (HIV-infection), GSE6740_2 (HIV-infection, disease state), GSE7146 (hyperinsulinaemic, does response), GSE7765 (dose response, DMSO or 100 nM Dioxin), GSE8441 (dietary intake response), GSE9574 (breast cancer), and GSE9499 (hypomorphic germline mutations). The diversity of these datasets will ensure that performance of the proposed algorithms is not due to specific characteristics of the data.

3.2 Bias of DEG identification algorithms

Statistics based DEG identification algorithms such as t-test and fold change all have different bias in their ranking statistics. There are three factors in their ranking criteria: $r(g) = (d, e, v)$ where d is the difference of expression levels between two conditions; e is the overall gene expression level of the gene; and v is the variance of gene expression. T-statistics based methods may make false positive prediction for genes with low d because of small v. Fold change method instead suffers from the fact that a gene with large variances tend to have larger fold changes. It is thus interesting to visualize the bias of different DEG algorithms in the (d, e) feature space. For simplicity, the variance v feature is neglected as it is not correlated to true DEGs as strong as (d, e) features.

In Fig1, we showed that most of true DEGs are located in the sparse boundary regions as outliers in the (d, e) space. A smaller portion of true DEGs are mixed with other non-DEGs in the dense regions and cannot be differentiated by the algorithms such as FC. To illustrate the bias of different DEG identification algorithms, for each algorithm, we showed in Fig. 2 true positive (TP), true negative (TN), false positive (FP), and false negative (FN) DEGs for dataset GSE9499 which has 77 DEGs. Fig.2. (a) shows that fold change method missed most true DEGs that are located in the region below the cutoff and with high expression levels (see FN genes). This is because FC uses a fixed ratio as cutoff. It also made many false positives most in the region with low expression levels (see FP genes). Rank product (Fig.2.(b)) missed similar true DEGs with fold change but the false positive genes have different distribution. Fig. 2 (c) shows the predicted DEGs of t-test. This method missed a lot of true DEGs that have high average difference between two conditions. Most of its false positives are located across the expression level with low average difference, reflecting the fact that it can be misled by genes with small variances.

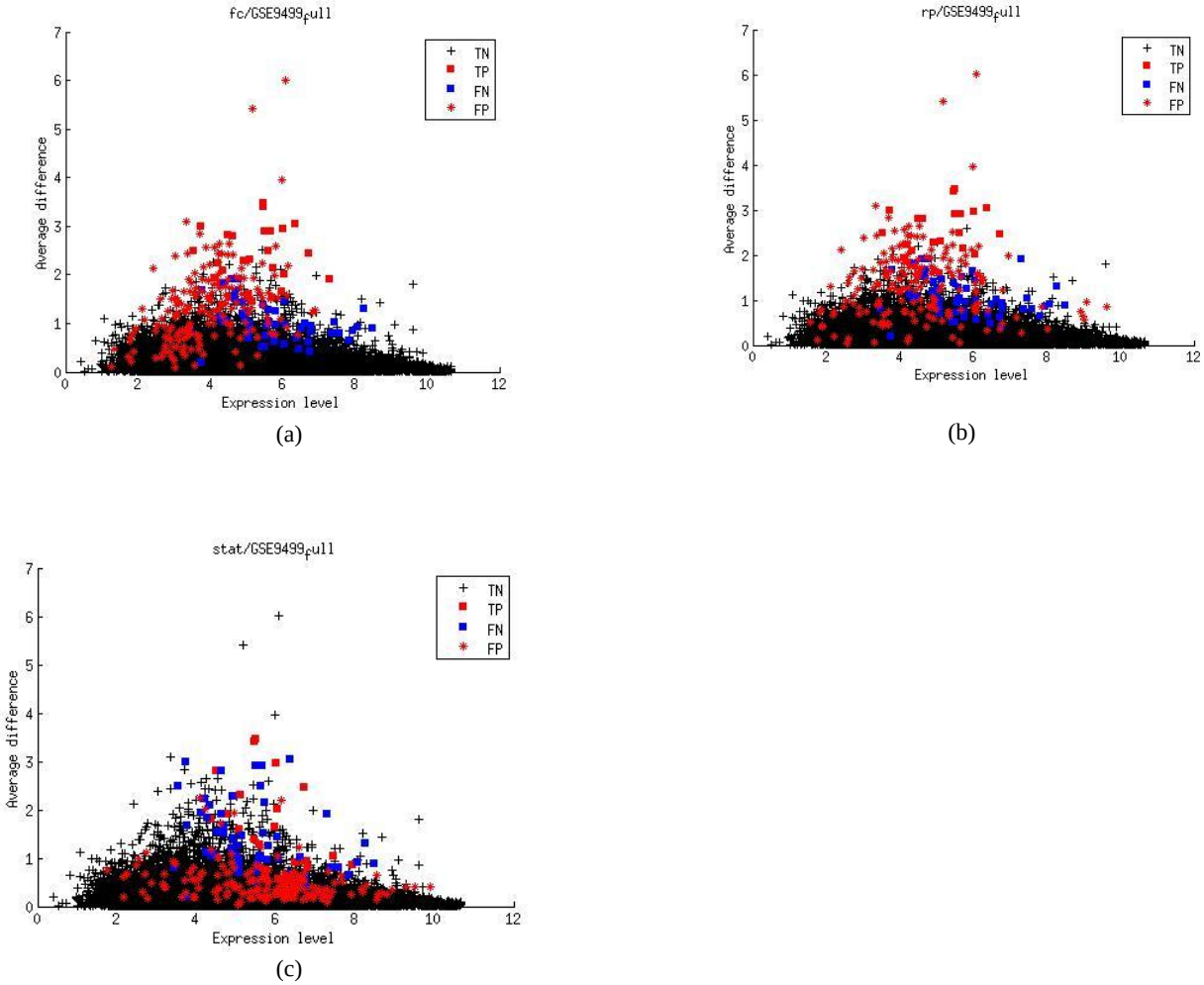


Fig.2. Bias of DEG identification algorithms: fold change (a), rank product (b), and t-test (c). Many of the false positive predictions are located in the dense regions, which can be easily removed using the proposed Pareto gene pruning algorithm.

3.3 Improving DEG identification algorithms using Pareto gene pruning

3.3.1 Effect of Pareto gene pruning of non-DEGs

To test the enrichment of true DEGs after pruning, we applied the Pareto gene pruning algorithm to the 17 microarray datasets each having 22283 genes. The idea is that by pruning those non-DEGs, true DEGs are more enriched in the remaining list and can be easier to be detected by other algorithms. The reduced search range of candidate DEGs can already greatly reduce computational cost for detecting combinatorial gene sets. By applying Pareto pruning to 17 datasets, it is found that this procedure can prune more than ≥81% genes out of 22283 while keeping ≥87% out of 184 true DEGs. The enrichment of true DEGs in the pruned gene sets will significantly improve their accuracy. Note that we provide a binary search procedure to determine the parameters for keeping a user-specified number of candidate genes.

3.3.3 Improving standard DEG algorithms using Pareto pruning

To evaluate the improvement of prediction performance of DEG algorithms with Pareto pruning, we use the receiver operating characteristic (ROC), or simply ROC curve. It is a graphical plot of the fraction of true positives (TPR = true positive rate) vs. the fraction of false positives (FPR = false positive rate) as the K (the number of genes predicted to be DEGs) varies. We use the area under curve (AUC) value of the ROC curve as the criterion for comparison. To make the comparison relevant to real-world practice, we only plot and compare the AUC value with K varies from 1 to 1000 rather than to K=22832 as done previously. The reason is that biologists rarely have the resources to check all 22832 genes and usually only care about few top predicted DEGs for experimental verification.

We calculate AUC values with K up to 1000 for four DEG algorithms with or without using Pareto gene pruning. The experiments are conducted on all 17 datasets with total 284 true DEGs. Table 1 shows that Pareto gene pruning significantly improved the AUC values for all four popular DEG algorithms especially for fold change with 18.4% improvement and rank product algorithm with 48.1% increase of AUC score. To get a more concrete intuition of the effect of Pareto gene pruning, Table 2 shows the total numbers of true DEGs out of top K predictions identified by different algorithms for the 17 datasets with or without Pareto gene pruning. First, the results showed that rank product and fold change have worse performance than the tTest algorithm in terms of identifying experimentally verified true DEGs. For example, tTest can detect 132 true DEGs from the 17 datasets when K=150 predictions are allowed for each dataset. Instead, RP and FC can only detect 74 and 97 true DEGs respectively. When the no. of predictions K increases, all algorithms cover more true DEGs with the highest coverage by t-test algorithm which retrieves 222 out of 284 true DEGs when 550 genes are allowed to predict for each dataset. A major observation of Table 2 is that all three algorithms can benefit from Pareto gene pruning with maximum benefit for tTest. In the case of RP, Pareto pruning helps RP to find 60 (nearly 56.6%) more true DEGs for K=550. For FC and tTest, 24% and 10% more true DEGs are identified with the help of Pareto gene pruning.

Table 1. Improvement of AUC values for DEG algorithms after Pareto pruning: Rp, Fc, and tTest.

	Partial AUC	Improvement
Rp/Rp'	0.0162/0.024	48.1%
Fc/Fc'	0.0245/0.029	18.4%
tTest/tTest'	0.0284/0.031	9.2%

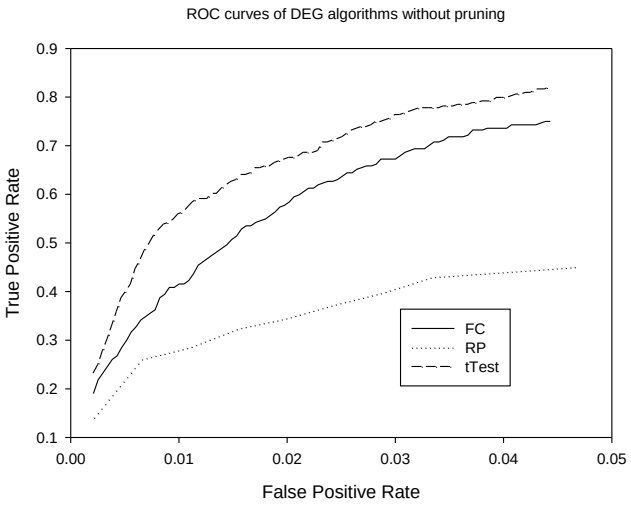
Table 2. Increase of No. of detected true DEGs out of top K predictions with or without Pareto pruning. Rp', tTest', FC' are algorithms with Pareto pruning. The total number of true DEGs of the 17 datasets is 284.

	K=150	K=250	K=350	K=450	K=550
Rp/Rp'	74/78	81/97	92/118	98/143	106/166
Fc/Fc'	97/106	120/143	146/177	164/203	178/221
tTest/tTest'	132/138	163/173	179/193	191/211	202/222

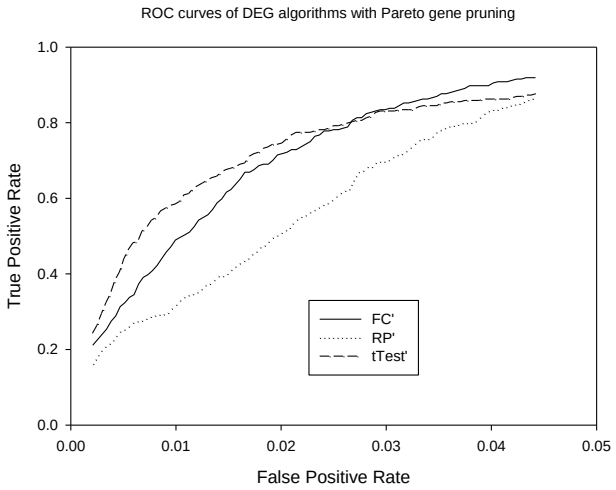
To further investigate the improvement of Pareto pruning over classic DEG algorithms, Fig 3(a)-(c) show the ROC curves of the algorithms with and without pruning with top K=1 to 1000. Fig 3(a) shows that t-test algorithm has the best performance and there exists dominance relationship of t-test>FC>RP. Fig 3(b) shows that after applying Pareto gene pruning, the performance of t-test, rank product, and fold change algorithms are all significantly improved. Fig 3.(c) shows the complete comparison of the three DEG identification algorithms with or without Pareto gene pruning. It clearly demonstrates the Pareto pruning has significantly improved the AUC values with improvements across all K ranging from 1 to 1000. Compared to the improvements of ROC curves as shown by the variance estimation algorithms, our improvements are much more significant.

4 Discussion

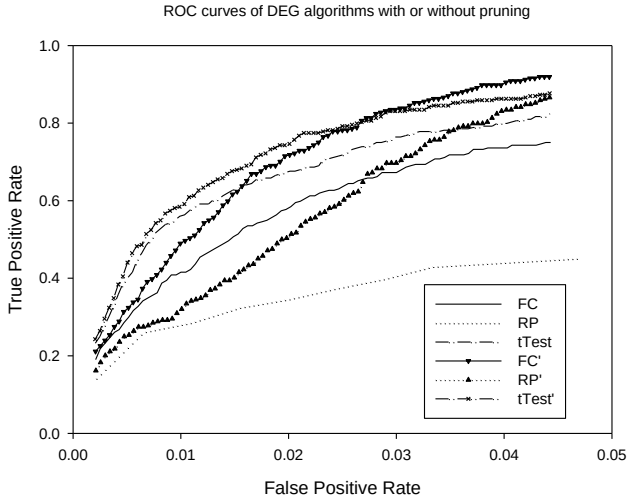
We have proposed a Pareto gene pruning algorithm that can prune non-differentially expressed genes with high confidence from the total gene list. This pruning procedure can significantly improve the prediction accuracy for popular DEG identification algorithms such as fold change, t-test, and rank product. The main benefit of Pareto gene pruning is not reducing running time of DEG identification algorithm, but significant improvement of prediction precision and recall as shown in the ROC curves. The improvement of DEG prediction performance comes from the observation that DEGs tend to have high expression values. There are several further improvements following this pattern recognition based DEG identification method. One is to use external datasets to estimate gene expression levels and difference of gene expressions. Our preliminary experiments showed estimating gene expression level is straightforward and feasible but estimating difference of gene expression needs more study. Another improvement is that additional features of DEGs can be introduced, e.g. the variance of gene expressions across multiple datasets. For example, the variance estimation method using multiple datasets [25] can be combined with our Pareto pruning algorithm. Functional annotation information from gene ontology or pathways can also be integrated to help the pruning process.



(a)



(b)



(c)

Fig.3. Comparison of ROC curves of DEG algorithms with (b) or without (a) Pareto pruning. The improvements of prediction accuracy can be clearly observed in (c). It shows dramatic improvement on the prediction performance of rank product as well as t-test and fold change with the Pareto gene pruning

ACKNOWLEDGEMENTS

This work is partially supported by National Science Foundation/EPSCoR under Grant No. EPS-0447660.

REFERENCES

[1] J. L. Derisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, no. 5338, pp. 680-686, Oct.1997.

[2] R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk, "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments," *FEBS Lett.*, vol. 573, no. 1-3, pp. 83-92, Aug.2004.

[3] I. B. Jeffery, D. G. Higgins, and A. C. Culhane, "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data," *Bmc Bioinformatics*, vol. 7 July2006.

[4] D. Nam and S. Y. Kim, "Gene-set approach for expression pattern analysis," *Briefings in Bioinformatics*, vol. 9, no. 3, pp. 189-197, May2008.

[5] J. Shi and M. G. Walker, "Gene set enrichment analysis (GSEA) for interpreting gene expression

profiles," *Current Bioinformatics*, vol. 2, no. 2, pp. 133-137, May2007.

[6] D. Nam and S. Y. Kim, "Gene-set approach for expression pattern analysis," *Briefings in Bioinformatics*, vol. 9, no. 3, pp. 189-197, May2008.

[7] R. Chen, A. A. Morgan, J. Dudley, T. Deshpande, L. Li, K. Kodama, A. P. Chiang, and A. J. Butte, "FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease," *Genome Biol.*, vol. 9, no. 12, p. R170, Dec.2008.

[8] L. Nevarez, V. Vasseur, D. G. Le, A. Tanguy, I. Guisle-Marsollier, R. Houlgatte, and G. Barbier, "Isolation and analysis of differentially expressed genes in Penicillium glabrum subjected to thermal stress," *Microbiology*, vol. 154, no. Pt 12, pp. 3752-3765, Dec.2008.

[9] M. Estler, G. Boskovic, J. Denvir, S. Miles, D. A. Primerano, and R. M. Niles, "Global analysis of gene expression changes during retinoic acid-induced growth arrest and differentiation of melanoma: comparison to differentially expressed genes in melanocytes vs melanoma," *Bmc Genomics*, vol. 9, p. 478, 2008.

[10] L. Satish, W. A. Laframboise, D. B. O'Gorman, S. Johnson, B. Janto, B. S. Gan, M. E. Baratz, F. Z. Hu, J. C. Post, G. D. Ehrlich, and S. Kathju, "Identification of differentially expressed genes in fibroblasts derived from patients with Dupuytren's Contracture," *BMC Med. Genomics*, vol. 1, p. 10, 2008.

[11] L. Shi, W. D. Jones, R. V. Jensen, S. C. Harris, R. G. Perkins, F. M. Goodsaid, L. Guo, L. J. Croner, C. Boysen, H. Fang, F. Qian, S. Amur, W. Bao, C. C. Barbacioru, V. Bertholet, X. M. Cao, T. M. Chu, P. J. Collins, X. H. Fan, F. W. Frueh, J. C. Fuscoe, X. Guo, J. Han, D. Herman, H. Hong, E. S. Kawasaki, Q. Z. Li, Y. Luo, Y. Ma, N. Mei, R. L. Peterson, R. K. Puri, R. Shippy, Z. Su, Y. A. Sun, H. Sun, B. Thorn, Y. Turpaz, C. Wang, S. J. Wang, J. A. Warrington, J. C. Willey, J. Wu, Q. Xie, L. Zhang, L. Zhang, S. Zhong, R. D. Wolfinger, and W. Tong, "The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies," *BMC Bioinformatics*, vol. 9 Suppl 9, p. S10, 2008.

[12] K. Yang, J. Li, and H. Gao, "The impact of sample imbalance on identifying differentially expressed genes," *BMC Bioinformatics*, vol. 7 Suppl 4, p. S8, 2006.

[13] P. Baldi and A. D. Long, "A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes," *Bioinformatics*, vol. 17, no. 6, pp. 509-519, June2001.

[14] N. Jain, J. Thatte, T. Braciale, K. Ley, M. O'Connell, and J. K. Lee, "Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays," *Bioinformatics*, vol. 19, no. 15, pp. 1945-1951, Oct.2003.

[15] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. Natl. Acad. Sci. U. S A*, vol. 98, no. 9, pp. 5116-5121, Apr.2001.

[16] A. Wille, W. Gruissem, P. Buhlmann, and L. Hennig, "EVE (external variance estimation) increases statistical power for detecting differentially expressed genes," *Plant Journal*, vol. 52, no. 3, pp. 561-569, Nov.2007.

[17] R. D. Kim and P. J. Park, "Improving identification of differentially expressed genes in microarray studies using information from public databases," *Genome Biology*, vol. 5, no. 9 2004.

[18] K. Kadota, Y. Nakai, and K. Shimizu, "A weighted average difference method for detecting differentially expressed genes from microarray data," *Algorithms for Molecular Biology*, vol. 3 June2008.

[19] M. Dettling, E. Gabrielson, and G. Parmigiani, "Searching for differentially expressed gene combinations," *Genome Biol.*, vol. 6, no. 10, p. R88, 2005.

[20] K. Kadota, Y. Nakai, and K. Shimizu, "A weighted average difference method for detecting differentially expressed genes from microarray data," *Algorithms for Molecular Biology*, vol. 3 June2008.

[21] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar, "NCBI GEO: mining tens of millions of expression profiles - database and tools update," *Nucleic Acids Research*, vol. 35, p. D760-D765, Jan.2007.

[22] R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk, "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments," *FEBS Lett.*, vol. 573, no. 1-3, pp. 83-92, Aug.2004.

[23] F. X. Hong, R. Breitling, C. W. McEntee, B. S. Wittner, J. L. Nemhauser, and J. Chory, "RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis," *Bioinformatics*, vol. 22, no. 22, pp. 2825-2827, Nov.2006.

[24] K. Kadota, Y. Nakai, and K. Shimizu, "A weighted average difference method for detecting differentially expressed genes from microarray data," *Algorithms for Molecular Biology*, vol. 3 June2008.

[25] A. Wille, W. Gruissem, P. Buhlmann, and L. Hennig, "EVE (external variance estimation) increases statistical power for detecting differentially expressed genes," *Plant Journal*, vol. 52, no. 3, pp. 561-569, Nov.2007.