BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 4 2005, pages 445–450
doi:10.1093/bioinformatics/bti189

Finding differentially expressed genes for pattern generation

Osman Abul[1,2], Reda Alhajj[1,*], Faruk Polat[2] and Ken Barker[1]

[1]Department of Computer Science, University of Calgary, Calgary, Alberta, Canada and [2]Department of Computer Engineering, Middle East Technical University, AnKara, Turkey

Received on May 21, 2004; accepted on November 11, 2004
Advance Access publication December 17, 2004

ABSTRACT

Motivation: It is important to consider finding differentially expressed genes in a dataset of microarray experiments for pattern generation.

Results: We developed two methods which are mainly based on the q-values approach; the first is a direct extension of the q-values approach, while the second uses two approaches: q-values and maximum-likelihood. We present two algorithms for the second method, one for error minimization and the other for confidence bounding. Also, we show how the method called *Patterns from Gene Expression* (PaGE) (Grant *et al*., 2000) can benefit from q-values. Finally, we conducted some experiments to demonstrate the effectiveness of the proposed methods; experimental results on a selected dataset (BRCA1 vs BRCA2 tumor types) are provided.

Contact: alhajj@cpsc.ucalgary.ca

1 INTRODUCTION

The advent of the microarray technology enabled researchers to measure the expression levels of thousands of genes in a single experiment. As a result, a huge amount of gene expression datasets are being produced. The challenge is finding methods to analyze them for different needs, and to interpret the obtained results for discoveries.

Using the microarray technology for gene expression measurement is a breakthrough. However, it has some inherent problems, including dimensionality, noise, and natural variability. In order to alleviate these deficiencies, expression levels of genes are generally measured in replicas. So, if there is large number of replicas for a specific gene in a specific condition, then we can use them to induce its distribution; but generally only few replicas are available.

The problem of identifying differentially expressed genes can be stated as follows: *given replicated gene expression measurements of two conditions (reference and control), find a subset of all genes having significant expression levels across these two conditions*. This definition implies that there are some other subsets where the change is not significant. Given the set of differentially expressed genes, we can further group them as over-expressed (up-regulated) and under-expressed (down-regulated) genes.

In case of a large number of replicas, the average gene expression may be estimated using the well-known statistics approaches. However, classical statistics approaches are very conservative when small number of replicas is available. For 2 and 3 replica experiments, differentially expressed genes are, respectively, $|\overline{x}_{g,i} - \overline{x}_{g,0}| > 22.3\widehat{\sigma}$ and $|\overline{x}_{g,i} - \overline{x}_{g,0}| > 5.2\widehat{\sigma}$, where $\widehat{\sigma}$ is the estimated standard error, $\overline{x}_{g,k}$ is the average expression of gene g in condition k, and the statistical significance level is 1% (Claverie, 1999). This motivated researchers to design new methods for finding differentially expressed genes; some sophisticated methods to handle this problem include PaGE (Grant *et al*., 2000) and q-values (Storey, 2002).

Although differentially expressed genes exhibit significant changes between two conditions, their significance levels are not the same. The *pattern generation* problem further clusters significant genes based on ordered significance levels.

In this paper, we present an approach which is mainly concerned with pattern generation using q-values in order to find differentially expressed genes in a dataset of microarray experiments. In particular, this paper proposes two methods; the first is a direct extension of the q-values approach, while the second uses q-values and maximum-likelihood. Also, we show how PaGE can benefit from q-values. Finally, we conducted some experiments to demonstrate the effectiveness of the proposed methods; experimental results achieved on a selected dataset (BRCA1 vs BRCA2 tumor types) are promising.

The rest of the paper is organized as follows. Pattern generation using PaGE is covered in Section 2. Our approach of pattern generation using q-values is presented in Section 3. Experimental results are given in Section 4. Discussion and conclusions are included in Section 5.

[*]To whom correspondence should be addressed.

2 PATTERN GENERATION USING PaGE

The problem of finding differentially expressed genes has recently received considerable attention. PaGE (Manduchi *et al*., 2000; Grant *et al*., 2000) handles the problem of confidence estimation for identifying differentially expressed genes in order to generate patterns. In the settings, there are several conditions for every gene, and for each condition there are replicas. The process works as follows. First, a reference and a control condition are selected. Second, average values of replicas are calculated. Third, the shifted average expression value of the control condition of each gene is divided by that of the reference condition. A gene for which the ratio is above a cutoff ratio ($Cr > 1$), is considered up-regulated in control condition compared to reference condition.

Consider a gene, say g, which is not up-regulated and let $\mu_{g,i}$, $\mu_{g,0}$, $X_{g,i}$ and $X_{g,0}$, respectively, be the mean value of distribution and the distribution of samples for control and reference conditions of gene g. Then, the probability of false-positives (FPR, a.k.a. Type I error) is:

$$Prob\left(\frac{\overline{X}_{g,i}}{\overline{X}_{g,0}} > Cr|\frac{\mu_{g,i}}{\mu_{g,0}} \leq 1\right) \tag{1}$$

Based on the fact that $\frac{\mu_{g,i}}{\mu_{g,0}} \leq 1$, an upper bound on Formula (1) can be obtained as follows:

$$Prob\left(\frac{\frac{\overline{X}_{g,i}}{\mu_{g,i}}}{\frac{\overline{X}_{g,0}}{\mu_{g,0}}} > Cr|\frac{\mu_{g,i}}{\mu_{g,0}} \leq 1\right) \tag{2}$$

It is required to find a least cutoff value of Cr, satisfying the maximum false-positive rate s (say 1%, for example). This is because any cutoff larger than Cr increases significance, and at the same time reduces power, i.e., Cr is the best balance of FPR and power. Using the fact that two events on both sides of the symbol "|" are independent in Formula (2), we get the following formula,

$$Prob\left(\frac{\frac{\overline{X}_{g,i}}{\mu_{g,i}}}{\frac{\overline{X}_{g,0}}{\mu_{g,0}}} > Cr\right) < s\% \tag{3}$$

In PaGE, the Bayes' formula can be used to determine the confidence level for each gene. Here, note that the conditional probability in Formula (4) is FPR and (1–*Confidence*) is the false-discovery rate. Also, since *Prob*(*not up*) is neither known nor estimated, the approximation is done based on the worst-case that *Prob*(*not up*) = 1.

$$\begin{aligned} &Confidence \\ &\quad = 1 - Prob(not\ up|predicted\ up) \\ &\quad = 1 - \frac{Prob(not\ up)Prob(predicted\ up|not\ up)}{Prob(Predicted\ up)} \\ &\quad \leq 1 - \frac{Prob(predicted\ up|not\ up)}{Prob(Predicted\ up)} \end{aligned} \tag{4}$$

Table 1. Multiple-hypothesis testing scenario of $M_{(-|-)}$ genes

	$H_g = 0$	$H_g = 1$	
retain $H_g = 0$	$M_{(0\|0)}$	$M_{(0\|1)}$	$M_{(0\|-)}$
reject $H_g = 0$	$M_{(1\|0)}$	$M_{(1\|1)}$	$M_{(1\|-)}$
	$M_{(-\|0)}$	$M_{(-\|1)}$	$M_{(-\|-)}$

It is clear that if exact value of *Prob*(*not up*) < 1 is known, then the power will be increased within the same confidence level. This means that large number of differentially expressed genes will be recalled. In the next section we show that this probability can be estimated from the dataset.

3 THE PROPOSED METHODS

In this section, we present our approach of using q-values for pattern generation. To the best of our knowledge, pattern generation from q-values has not been studied thoroughly before. We first present how p-values and q-values are computed. Then, we introduce the two proposed methods to be used for pattern generation.

3.1 Computing p-values and q-values

Let $H_g = 0$ be the null hypothesis; and let $H_g = 1$ be the alternative hypothesis. Under the one-tailed null hypothesis assumption, one minus the probability of rejecting null hypothesis against an observation ($X = x$) is known as its p-value, i.e., $p - value(x) = 1 - P(X < x|H_g = 0)$. The smaller the p-value, the more evidence it is against the null hypothesis. For single hypothesis testing, if p-value of the observation $X = x$ is less than a predefined significance value ($0 < \alpha < 1$), then the null hypothesis is rejected, otherwise it is retained.

There are usually thousands of features (genes) in microarray experiments, and this necessitates considering thousands of tests at the same time. Therefore, we should adopt multiple-hypothesis testing rather than single-hypothesis testing. In this regard, there are a number of schemes proposed for multiple-hypothesis testing, e.g., *Bonferroni*, *Fixed threshold*, and *First r*; they all differ in conservation levels. Table 1 characterizes multiple-hypothesis testing scenario of $M_{(-|-)}$ number of genes. The number of false-positives (Type I error), false-negatives (Type II error), true-negatives, and true-positives are $M_{(1|0)}, M_{(0|1)}, M_{(0|0)}, M_{(1|1)}$, respectively. Also, $M_{(0|-)}$ and $M_{(1|-)}$ are, respectively, predicted number of non-differentially and differentially expressed genes; while, $M_{(-|0)}$ and $M_{(-|1)}$ are, respectively, true number of non-differentially and differentially expressed genes; and $\frac{M_{(1|1)}}{M_{(-|1)}}$ gives the *power* of the test.

So, given the replicas of control and reference samples, it is possible to compute the t-statistics for each gene by using the following formula under the assumption that genes have differing standard deviations (Dudoit *et al.*, 2000),

$$t_g = \frac{\overline{x}_{g,i} - \overline{x}_{g,0}}{\sqrt{\frac{s_{g,i}^2}{n_i} + \frac{s_{g,0}^2}{n_0}}} \qquad (5)$$

where $\overline{x}_{g,i}$ and $\overline{x}_{g,0}$ are means of replicas of control and reference conditions with respective standard deviations $s_{g,i}^2$ and $s_{g,0}^2$, and replica counts n_i and n_0 for gene g. It is clear that t-statistics favor for large mean differences and small standard deviations.

In case that we do not know the exact distribution, the p-value for each gene can be computed by using the *Permutation algorithm.* So, t-statistics of each gene is computed for each permutation. From these statistics, the p-value of gene g can be computed using the following formula (Ge *et al.*, 2003).

$$p_g = \frac{\#\{b : |t_{g,b}| \geq |t_g|, b = 1 \cdots B\}}{B} \qquad (6)$$

where B is the number of permutations generated, t_g is the t-statistics for the original sample for gene g and $t_{g,b}$ is the t-statistics of $\acute{b}$th permutation of gene g. Alternatively, the formula below can be used for p-value computation (Storey, 2002).

$$p_g = \sum_{b=1}^{B} \frac{\#\{j : |t_{j,b}| \geq |t_g|, j = 1 \cdots M_{(-|-)}\}}{B \cdot M_{(-|-)}} \qquad (7)$$

Given the p-values for genes, the work described by Storey (2002) gives a method to compute q-values using p-values. The q-value for gene g is defined as the expected rate of false-positives in the set of genes having p-values smaller than or equal to that of g.

Computation of q-values are based on the *false discovery rate* concept, which is defined as the rate of false-positives in a given rejection region, i.e., $FDR = \frac{M_{(1|0)}}{M_{(1|-)}}$. By conditioning FDR on the case that the rejection region contains at least one gene, the author defined pFDR (positive FDR) as $pFDR = E\left(\frac{M_{(1|0)}}{M_{(1|-)}} | M_{(1|-)} > 0\right)$ The q-value of a p-value p_i is defined as $q(p_i) = min_{t \geq p_i} \widehat{FDR}(t)$; for a fixed rejection region $\widehat{FDR}(t)$ is defined as $\widehat{FDR}(t) = \frac{\widehat{\pi}_0 \cdot M_{(-|-)} \cdot t}{\#\{p_i \leq t\}}$, where the only unknown item in the formula is $\widehat{\pi}_0$, i.e., an estimate of $\pi_0 = \frac{M_{(-|0)}}{M_{(-|-)}}$. This estimate is given as follows, depending on the parameter $0 \leq \lambda < 1$.

$$\widehat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{M_{(-|-)}(1 - \lambda)} \qquad (8)$$

The above estimation can be evaluated for a given value of λ; and the method by Storey (2002) finds an optimum value for λ and thus for $\widehat{\pi}_0$. This exploits the fact that genes in the region $[\lambda..1]$ are observed to be uniformly distributed. So, we can easily compute q-values for all genes by estimating π_0 and having p-values computed for each gene.

One advantage of q-value over p-value is that it does not force the selection of a particular rejection region. Another advantage of q-value is related to the fact that by bounding false-discoveries, the amount of waste of time and money can also be bounded with the same rate of false-discoveries beforehand. This is because significantly selected genes tend to experimentally verify in microarray experiments.

3.2 Extending the q-values approach

This method uses the information on estimated number of differentially expressed genes, which is reasonable to find exactly the intended number of genes. After finding this set of genes, pattern values can be associated to each gene based on their experimental averages. The genes in the estimated set of non-differentially expressed genes can be assigned pattern 0, meaning that these genes show insignificant expression differences between the two conditions.

As it has already been shown above, π_0 can be estimated from p-values. This estimation gives the proportion of non-differentially expressed genes. So, $1-\pi_0$ gives the proportion of differentially expressed genes. Using the estimator of π_0, estimated values of $M_{(-|0)}$ and $M_{(-|1)}$ can be found easily; and from the sorted q-values, we can find an index *ind* as:

$$Min_i \left\{ i | (i - (q\text{-}value(i) \times i)) \geq M_{(-|1)}, \right.$$
$$\left. i = M_{(-|1)}, \ldots, M_{(-|-)} \right\} \qquad (9)$$

where $q\text{-}value(i) \times i$ gives the number of non-differentially expressed genes in the region [1..*i*]. So, $i - q\text{-}value(i) \times i$ gives the number of differentially expressed genes in the same region.

The region between [1..*ind*] on sorted q-values contains all the differentially expressed genes with a high confidence; and hence the region $[ind + 1..M_{(-|-)}]$ contains mainly non-differentially expressed genes. So, these genes do not discriminate much between control and reference samples, and can be considered as non-discriminating genes. For this reason, these genes can be omitted from the dataset for feature extraction purposes. Since those genes are not differentially expressed, we can attribute them as having zero differential expression pattern values. The region $[1..ind]$ contains both differentially expressed and non-differentially expressed genes; further investigation is required to assign patterns to them. The additional information here is that among *ind* number of genes, the estimated number of significant genes is $M_{(-|1)}$.

For further investigation, we cannot apply the q-values approach again for the region [1..*ind*] because this time, as we have additional information, it is not valid to assume $H_g = 0$. So, we can make use of the mean cutoff ratio method (as in PaGE) for finding up-regulated and down-regulated genes. By exploiting these facts, we can sort the region based on the absolute value of the *log* transformed mean ratios, and then select the up-regulation cutoff (Cr) and the down-regulation cutoff (cr). After sorting, we can find the values of *Cr* and *cr* as follows:

$$Cr = Min_i \left(\frac{\overline{x}_{g,i}}{\overline{x}_{g,0}} | \frac{\overline{x}_{g,i}}{\overline{x}_{g,0}} > 1, \quad i = M_{(-|1)} + 1, \ldots, ind \right) \tag{10}$$

$$cr = Min_i \left(\frac{\overline{x}_{g,i}}{\overline{x}_{g,0}} | \frac{\overline{x}_{g,i}}{\overline{x}_{g,0}} < 1, \quad i = M_{(-|1)} + 1, \ldots, ind \right) \tag{11}$$

With this scheme, all the genes in the sorted region $[1..M_{(-|1)}]$ are assigned either positive or negative pattern values. Within this region, pattern values are computed using powers of *Cr* and *cr*. The genes in the region $[M_{(-|1)} + 1..ind]$ are assigned zero pattern values. So, there are exactly $M_{(-|1)}$ number of genes assigned either positive or negative patterns.

To determine differentially expressed genes in the region [l..*ind*], the PaGE approach can also be used as an alternative to the method presented above. The problem with the PaGE approach is that it does not use the prior information of the estimated number of significant genes. However, we can extend PaGE to use this information to increase its power.

3.3 Using q-values and maximum likelihood

The first $M(-|1)$ genes in the list sorted in ascending order can be assumed to be differentially expressed genes; they form the maximum-likelihood set. So, the main question is how to partition a set of differentially expressed genes into up-regulated and down-regulated classes. The natural answer is to decide by looking at their sample mean expression ratios; if the control to reference ratio is above 1, then we can consider the gene as up-regulated, otherwise it is down-regulated.

By using Bayes's rule, we can obtain the following formulas for the true-positive rate (TPR), the false-negative rate (FNR), the true-negative rate (TNR), Confidence, FDR, and the error rate (ER), depending only on FPR and the two *priors*, *Prob*(*predicted up*) and *Prob*(*not up*) of the up-regulation case.

$$\begin{aligned}
FPR &= Prob(predicted\ up|not\ up)\\
TPR &= \frac{Prob(predicted\ up) - Prob(not\ up)(FPR)}{1 - Prob(not\ up)}\\
FNR &= 1 - TPR\\
TNR &= 1 - FPR\\
Confidence &= 1 - \frac{Prob(not\ up)FPR}{Prob(predicted\ up)}\\
FDR &= 1 - Confidence\\
ER &= FNR + FPR\\
Accuracy &= 1 - ER
\end{aligned} \tag{12}$$

Similar formulas can be obtained for the down-regulation case by changing Formula (12) to express FPR as $FPR = Prob(predicted\ down|not\ down)$, and changing the priors in the other formulas by replacing *Prob*(*not up*) by *Prob*(*not down*) and *Prob*(*predicted up*) by *Prob*(*predicted down*).

To compute FPR, TPR, FNR, TNR, ER, Confidence, FDR, and Accuracy for given values of *Cr* for the up-regulation case, we need the two values, *Prob*(*not up*) and *Prob*(*predicted up*). The latter term can be directly computed because we have assumed *Cr* as given. But, the former term is not known; it can be estimated. Recall how we estimated up-regulated genes; similar arguments are valid for the down-regulation case.

Since every gene is attributed with one of the three characteristics: up-regulated, down-regulated or non-differentially expressed, then we can approach the problem from a classification perspective. We can then search for maximum-likelihood value of *Cr* and *cr* that best explains this labeling. From the determined values of *Cr* and *cr*, we can generate patterns as we did in the first method. Algorithm 3.1 is used to search the estimation of the cutoffs that minimize the total classification errors. Algorithm 3.1 uses the fact that *Cr* is bounded up with maximum ratio in the dataset and down with 1; the case of *cr* is similar.

Algorithm 3.1 Error Minimization

```
1. Compute $p$-values for all genes under the null hypothesis;
2. Determine cutoff values for $Cr^{max}$ and $cr^{min}$, simply by finding the maximum and minimum ratios, respectively;
3. Estimate the true values for $M_{(-|_0)}$ and $M_{(-|1)}$, using the $q$-values approach;
4. Discretize the interval $[cr^{min}..1]$ into $I1$;
5. Discretize the interval $[1..Cr^{max}]$ into $I2$;
6. For each value of $(I1, I2)$ pairs:
    - Find the corresponding values of Cr and cr;
    - Compute the error rate value for up-regulation $(ER1)$ and down-regulation $(ER2)$;
    - Find the value of $Min(ER1 + ER2)$ and return the corresponding value of Cr and cr;
```

For each interval-value pair (an instantiation of *Cr* and cr), the error rate is computed separately for up-regulation and down-regulation, and the cutoff values are selected in such a way that gives the minimum total error. Note that instead of error minimization, we can also use accuracy maximization; both will give the same result.

The basic idea developed here can also be applied to generate patterns based on confidence bounding, i.e., patterns having confidence above a given confidence value. For confidence computation, we use the respective part in Formula (12). The basic algorithm is similar to Algorithm 3.1; but they only differ in operations of the last step. More precisely, the last step should be replaced with the snippet given below.

```
1. For each value of $I1$ (considered in decreasing $cr$ values)
   - Find the corresponding $cr$;
   - Compute the Confidence using down-regulation variant of Formula (12);
   - If the computed confidence is less than the user-specified value then continue
     else break with the current $cr$;
2. For each value of $I2$ (considered in increasing $Cr$ values)
   - Find the corresponding $Cr$;
   - Compute the Confidence using Formula (12);
   - If the computed confidence is less than the user-specified value then continue
     else break with the current $Cr$;
```

During the search, if the current confidence is lower than the user specified minimum value, then the search continues; otherwise it ends and the current cutoff is returned.

For both error minimization and confidence satisfaction variants, there are three alternative sets of genes to consider for pattern generation. The first includes all genes, the second is a selected subset of $M_{(-|1)}$ genes, and the third contains the sorted first *ind* genes (recall the first method). The first is the default in our implementation.

4 EXPERIMENTAL RESULTS

The dataset used in our experiments is a normalized breast cancer data from (Hedenfalk *et al*., 2001). We have used BRCAl as the reference and BRCA2 as the control condition. There are 7 replicas for BRCAl and 8 for BRCA2. Finally, we have used 3170 genes out of 3226 as suggested by Storey (2002), because the remaining genes contain mainly outliers.

In our implementation, we have tried both Equations (6) and (7) for computing the p-values. Equation (6) is computationally more efficient and allows all of the combinations, $\binom{15}{7} = 6435$, to be tried in a reasonable time. Another advantage is that it allows the user to get the same results in each run, since there is no randomness in the process. On the other hand, PaGE gives different number of expressed genes and levels for each run, even without changing the parameters or the dataset; simply because PaGE has a random component. For q-value computation from p-values, we have used the script by Storey (2002).

4.1 Results of the First Method

The maximum ratio of differentially expressed genes has been found as 0.29388. So, at least $M_{(-|1)} = 931$ genes are differentially expressed between these two tumor types. For pattern generation, the *ind* value has been computed as 1722. So, it has been found that 1448 genes should be eliminated from the dataset. These genes are assigned zero pattern value. Among the 931 differentially expressed genes, 345 are up-regulated (patterns range from 1 to 4) and 586 are down-regulated (patterns range from -1 to -6). The cut-off ratios are 1.37 and 0.73 for *Cr* and *cr*, respectively. Finally, the q-value analysis results for the first 600 genes are illustrated in Figure 1. For instance, we expect that approximately 30 out of the first 300 genes are false-discoveries.

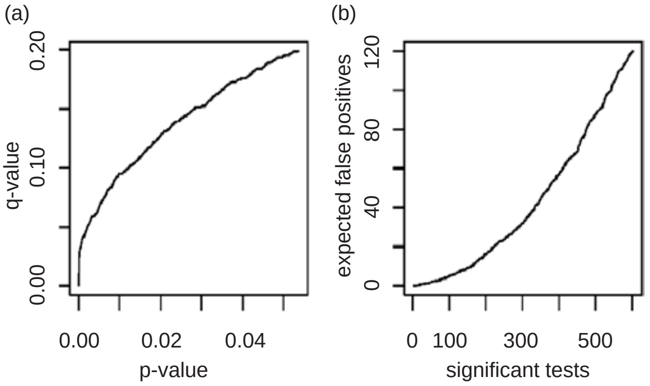

Fig. 1. BRCA2 vs. BRCA1 for the first 600 genes, (a) p-values vs. q-values (b) False discovery counts.

We have also tested the same dataset using PaGE with the confidence level of 0.6. It has been found that 126 are up-regulated and 356 are down-regulated genes, a total of 482. We have performed a second experiment using PaGE with the same confidence level of 0.6, and using a new dataset which has been formed by eliminating 1448 genes from the q-value analysis. It discovered 158 as up-regulated and 628 as down-regulated genes, a total of 786. Although, the dataset size has been reduced, PaGE discovered a larger number of significant genes. This is not a surprising result because removing a portion of the non-differentially expressed genes at the same significance level reduces the cutoff ratio. The argument here is that PaGE with reduced dataset has more power than PaGE with the initial dataset (786 vs. 482). Finally, the pattern ranges found with PaGE are $[-6..4]$ and $[-4..2]$ for 786 and 482 significant cases, respectively.

4.2 Results of the Second Method

The utilized dataset has maximum mean ratio of 4.34 and minimum mean ratio of 0.15. In the error minimization variant of the second method, we have found 657 up-regulated and 852 down-regulated genes with $Cr = 1.20$ and $cr = 0.78$. For the up-regulation case, the values found for FPR, FNR, and TPR are 0.12, 0.02, and 0.98, respectively. For the down-regulation case, respective values are 0.11, 0.06, and 0.94. The cutoffs pattern range is $[-7..8]$, and the total number of significant genes found is 1509. This is relatively higher than the estimated value (which is 931). But, this is not a contradiction because genes are found significant from classification point of view. For a correct understanding and interpretation of this result, consider the cardinality of the set that contains all significant genes found in the first method (which is 1722); 1509 and 1722 genes are suggested by feature selection for classification purposes by two different methods, and we interpret the result as consistent and provided extra information of eliminating $1722 - 1509 = 213$ genes. On the other hand, since these numbers are close, the two proposed methods verify each other. Further dimension reductions may be required for some applications. In this case, we consider that genes having low pattern values can be considered for removal.

We report two experiments with the confidence variant of the second method. The first experiment is done with minimum confidence of 80%. In this case, the cutoff values are $Cr = 1.54, cr = 0.73$, 220 up-regulated and 594 down-regulated genes; the pattern range is $[-6..3]$. In the second experiment, we have raised the minimum confidence to 90% and found the following results, $Cr = 1.90, cr = 0.64$, 88 up-regulated and 308 down-regulated genes; the pattern range is $[-4..2]$. As expected, the number of differentially expressed genes is reduced with higher minimum confidence. Finally, interval values of 100 are used for all experiments of the second method.

In all our experiments for both methods, we have found that the number of down-regulated genes is larger than the number of up-regulated genes. These results are consistent with the results reported by Hedenfalk *et al*. (2001) and Storey (2002).

5 DISCUSSION AND CONCLUSIONS

We have shown that by integrating the q-values approach with PaGE, it becomes possible to automate this process by making appropriate decreases or increases to suit the estimated true rates, of course if interest is in discovering approximately $M_{(-|1)}$ number of genes. Since we know the estimated number of differentially expressed genes, it is also possible to make some queries like: give me a set that contains at least half of the differentially expressed genes. With the knowledge of $M_{(-|1)}$, this kind of queries are possible to answer with PaGE, and can be automated. These are also valid for the confidence version of our second method. Finally, more important contribution of estimating the number of differentially expressed genes to PaGE is increasing the power in the same confidence level. As a future work, we aim to extend PaGE in that direction.

Our two methods make use of the q-values and the estimated number of significant genes in a way to assign patterns. The first method is dedicated for pattern generation using q-values. It eliminates non-discriminating genes from the dataset and reduces its size. This means that the q-values approach can also be used for feature selection purposes. Using the reduced dataset, the first method assigns genes' patterns using the mean expression cutoff approach, respecting the estimated number of differentially expressed genes.

In the second method, the q-values approach is only used to find the number of the true estimates of differentially expressed genes. Based on this value, a subset having cardinality equal to this number is selected as differentially expressed genes. This does not mean that all of the selected genes are more significant than all the other genes outside this set. The argument here is that the selected subset is the best maximum-likelihood of the sets having the same cardinality. Two algorithms have been developed based on this: one for error rate minimization and the other for confidence satisfaction. The two algorithms have been tested to demonstrate the effectiveness of the proposed methods; the results have been reported. Finally, PaGE has also been evaluated on both the initial dataset and the reduced dataset. As expected, PaGE with the reduced dataset has more power than PaGE with the initial dataset, for the same confidence level. This indicates that PaGE can benefit from the q-values approach. As another future work, we are planning to experiment these methods on other publicly available datasets.

REFERENCES

Claverie,J.M. (1999) Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet.*, **8**, 1821–1832.

Manduchi,E., Grant,G.R., McKenzie,S.E., Overton,G.C., Surrey,S. and Stoeckert,C.J.Jr. (2000) Generation of patterns from gene expression data by assigning confidence to differentially expressed genes. *Bioinformatics*, **16**, 685–698.

Dudoit,S., Yang,Y.H., Speed,T.P. and Callow,M.J. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Technical Report No. 578*, University of California, Berkeley, CA.

Ge,Y., Dudoit,S. and Speed,P.S. (2003) Statistical methods for identifying-differentially expressed genes in replicated cDNA microarray experiment. *Technical Report No. 633*, University of California, Berkeley, CA.

Storey,J.D. (2002) False discovery rates: theory and applications to DNA microarrays. PhD Thesis, Department of Statistics, Stanford University, CA.

Hedenfalk,I., Duggan,D., Chen,Y., Radmacher,M., Bittner,M., Simon,R., Meltzer,P. Gusterson,B., Esteller,M., Kallioniemi,O.P. *et al*. (2001). Gene-expression profiles in hereditary breast Cancer. *N. Engl. J. Med.*, **344**, 539–548.

Grant,G.R., Manduchi,E. and Stoeckert,C.J. (2000). Using non-parametric methods in the context of multiple testing to identify differentially expressed genes. *In Proceedings of Critical Assessment of Techniques for Microarray Data Analysis (CAMDA'00)*, Duke University, Durham, NC, December 18–19.