Gene selection for oligonucleotide array: an approach using PM probe level data

Dung-Tsa Chen1,∗, Sue-Hwa Lin2 and Seng-jaw Soong1

1Biostatistics and Bioinformatics Unit, Comprehensive Cancer Center, University of Alabama at Birmingham, 153 Wallace Tumor Institute, 1824 6th Avenue South, Birmingham, AL 35294, USA and 2Department of Molecular Pathology, University of Texas, M. D. Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030, USA

Received on May 3, 2003; revised July 15, 2003; accepted on September 10, 2003
Advance Access publication January 29, 2004

ABSTRACT

Motivation: Analysis of oligonucleotide array data, especially to select genes of interest, is a highly challenging task because of the large volume of information and various experimental factors. Moreover, interaction effect (i.e. expression changes depend on probe effects) complicates the analysis because current methods often use an additive model to analyze data. We propose an approach to address these issues with the aim of producing a more reliable selection of differentially expressed genes. The approach uses the rank for normalization, employs the percentile-range to measure expression variation, and applies various filters to monitor expression changes.

Results: We compare our approach with MAS and Dchip models. A data set from an angiogenesis study is used for illustration. Results show that our approach performs better than other methods either in identification of the positive control gene or in PCR confirmatory tests. In addition, the invariant set of genes in our approach provides an efficient way for normalization.

Contact: dtchen@uab.edu

INTRODUCTION

The oligonucleotide gene chip technique is powerful for analyzing mRNA expression (Yun et al., 2003; Yeatman, 2003; Winzeler et al., 2003; Takahashi et al., 2003). The technique allows researchers to investigate thousands of genes simultaneously and quickly, and therefore dramatically reduces labor, time, and cost involved in such analysis. Although the technology is a useful tool for monitoring mRNA expression, various questions have been raised including the use of probe measurement, normalization and modeling. In this paper, we address these issues by using oligonucleotide arrays from the Affymetrix GeneChip system.

An oligonucleotide array usually uses a set of oligo probe pairs to interrogate a target gene. Each probe pair consists of an oligonucleotide complementary to a portion of the target transcript (called perfect match or PM probe) and a companion oligonucleotide with the same sequence except with a single nucleotide mutation at the central position (called mismatch or MM probe). The companion oligonucleotide (i.e. the MM probe) is used to control for non-specific hybridization. Typically, a set of the PM and MM pairs are used for analysis. However, the use of MM as internal control has been questioned because about one-third of probes are found with MM>PM and there exists a high correlation between PM and MM (Irizarry et al., 2003b; Naef et al., 2002). Recently, several approaches have been proposed with the focus on PM intensity, such as Dchip (Li and Wong, 2001a; Zhou and Abagyan, 2002) and Affymetrix Microarray Suite version 5 (MAS 5.0) (Affymetrix, 2002).

Intensity, a surrogate of mRNA expression levels, may be influenced by various experimental factors, such as tissue handling, RNA quality, labeling efficiency, selection of reagents, and image scanning (Schadt et al., 2002; Brazma et al., 2002; Model et al., 2002). Normalization has been used to try to correct biases caused by these factors so that the adjusted intensities between arrays are on a common level and meaningful comparisons of gene expressions can be made. At least two types of normalization approaches exist: global and local normalization. However, each method has its own limitations. For example, the global normalization assumes that all genes are non-differentially expressed, and that experimental and unknown factors cause systematic change among arrays. Thus, the whole data set is used to estimate a normalization curve (Workman et al., 2002; Colantuoni et al., 2002; Hoffmann et al., 2002). The method is good at its efficiency to reduce variability among arrays, but has the risk of destroying biological patterns. In contrast, local normalization uses only a small portion of genes to assess the magnitude of system biases, such as housekeeping genes and spike genes (Schena...
Gene selection for oligonucleotide array data

et al., 1996; Wu, 2001; Hill et al., 2001). However, these genes have been shown to either have their intensity levels changed with experimental conditions of interest (Li and Wong, 2001b) or the intensity pattern was shown to be independent of the system biases (Schadt et al., 2002).

Several model-based approaches have been proposed to address the modeling issue. MAS 4.0 uses PM–MM or PM/MM, after deletion of probe outliers, to estimate a gene signal. This approach has been criticized for inefficient management of the MM data (Irizarry et al., 2003b). MAS 5.0 uses a log transformation of PM–MM to analyze the signals with which MM values are imputed when MM values are larger than PM values. The Tukey’s Biweight (TB) approach is then used to calculate a weighted mean for the gene signal. Intensities estimated from both MAS approaches are a summary statistic that has been simplified from the probe level data to the probe set level data. These approaches dramatically reduce the volume of data, from millions to thousands of data points, and render the data more manageable. However, their implementation is based on an assumption of homogeneity of probe affinity. The hypothesis may not be realistic because it has been shown that probe variations are often larger than chip variations (Li and Wong, 2001a; Irizarry et al., 2003a).

Dchip software analyzes probe level intensities using a multiplicative model to decompose each probe signal into a product of a gene expression index and probe-sensitivity index (Li and Wong, 2001b). The probe signals can be either PM–MM or PM only in the version 1.1 of the Dchip software. This approach has been successfully used in various genomic studies because of its efficient management of probe effect. However, this approach assumes that treatment and probe affinity effects enter the model in an additive way after the log transformation. The assumption postulates that a gene expression index is identical with respect to a probe set. When probe effects and treatment effects commingle, the assumption will be violated. For example, alternative splicing represents one form of interaction. Its occurrence often results in a gene with some sequences truncated or modified. It has been suggested that alternative splicing is common, occurring in 40% of human genes (Modrek and Lee, 2002) and produces multiple functional forms of many genes. Cross-hybridization also causes interaction. By cross-hybridization, we mean the binding of a probe to a non-target DNA sequence because of some shared similarity in the nucleotide sequences between the probe and the non-target DNA. This commonly occurs in microarray experiments. Thus, methods without consideration of interaction will likely either miss the target gene or identify false positive genes.

In the following sections, we propose an approach to addressing these issues. Data from a gene-chip study on the search for target genes of an antiangiogenic agent is explained in the Data section. Our approach to analyze the data is described in the evaluation section. Results and discussion are in the last two sections.

**METHODS**

We propose a new approach aiming at producing a more reliable selection of differentially expressed genes in the oligonucleotide array data. The approach uses the rank method to normalize PM intensities. Analyses of PM ranks are implemented by a percentile-range approach to measure probe variation and by two filters to select differentially expressed genes and invariant genes. The details of our approach are described as follows.

Rank has recently been used as a normalization tool in microarray data analysis (Kroll and Wolff, 2002; Hoyle et al., 2002), but the use is limited to the gene level data. Here we extend its application to the probe level data in oligonucleotide arrays. We hypothesize that the majority of gene expression levels are non-differential from one experimental condition to the other. Holding this assumption, it is likely to identify differentially expressed genes by using the variation of their ranks. The use of rank has several advantages over intensity in the analysis of oligonucleotide arrays. Rank provides a better treatment to alleviate effects of extreme values and to avoid assumption of any parametric distribution. Our experiences from various microarray data analyses have found that PM intensity in the oligonucleotide array data often shows a skewed distribution with a small population (5%) of extreme high intensities. Variation of PM is substantially large in this high intensity area compared to the areas of the low/moderate intensity levels. So the use of intensity tends to identify high abundance genes, which are often not the primary genes of research interest. In contrast, rank is more robust when it comes to outliers. The ranks of these extreme values are less influential than their intensity levels. In addition, unlike other measurements generated by complicated normalization, rank is a simple measure extracting order from intensity. The biomedical community can easily accept the concept. Also, because of highly positive correlation between intensity and its rank, the rank is likely to preserve most information of interest without distorting biological variation.

For analysis of rank, we first standardize PM rank (i.e. divided by the total number of probes in an array) to become a value ranged from 0 to 1. The value, denoted by $P$-rank, makes it comparable to parameter settings for various types of gene chips because each type of gene chip contains a different number of genes. Then we use a percentile-range to examine variation of $P$-rank. The percentile–range is defined as the difference between the $A$-th percentile and the $B$-th percentile of the $P$-ranks in a probe, where $0 \leq B < A \leq 100$. $A$ and $B$ are the upper and lower bounds of outlier thresholds, respectively. For studies using a few samples, such as in vitro studies, because the biological variation is well controlled, the two parameters $A$ and $B$ can be set up as 100 and 0, respectively. In this case, the percentile–range becomes the standard range. For studies involving large sample sizes, $A$ and $B$ can be 70–80 and 20–30, respectively, to eliminate outliers. This truncation
approach is similar to the trimmed means approaches, but, instead of estimation of means, it focuses on the variation using the range approach.

Three quantities of probe variation are calculated using the percentile-range: $D_{\text{range}}$, $D_{g1,g2}$ and $D_{g2,g1}$. $D_{\text{range}}$ has the same definition of the percentile-range whereas $D_{Ga,Gb}$ (e.g. $Ga$ and $Gb$ can be $g1$ and $g2$, respectively) is modification of the percentile-range. Specifically, $D_{Ga,Gb}$ is defined as the difference between the $B$-th percentile of the $P$-ranks in group $Ga$ and the $A$-th percentile of the $P$-ranks in group $Gb$ in a probe. The mathematical formulae can be expressed as follows.

Let $Y_{i,j,k}$ be a $P$-rank for the $j$-th probe in the $i$-th gene of array $k$. Assume group $Ga$ has arrays $a1, \ldots, a_n$ and group $Gb$ has arrays $b1, \ldots, b_n$. Then,

$$D_{\text{range}} = A \text{-th percentile of } \{Y_{i,j,b1}, \ldots, Y_{i,j,b_n}\}$$

$$- B \text{-th percentile of } \{Y_{i,j,a1}, \ldots, Y_{i,j,a_n}\}$$

$$D_{Ga,Gb} = B \text{-th percentile of } \{Y_{i,j,a1}, \ldots, Y_{i,j,a_n}\}$$

$$- A \text{-th percentile of } \{Y_{i,j,b1}, \ldots, Y_{i,j,b_n}\}$$

Here we use $D_{\text{range}}$ to examine the maximum variation among all arrays for a probe. $D_{\text{range}}$ measures the overall variation and therefore allows us to select invariant genes. In contrast, $D_{g1,g2}$ and $D_{g2,g1}$ are used to evaluate the minimum difference between two groups of arrays for a probe (i.e. between group variation). A large positive value of $D_{g1,g2}$ or $D_{g2,g1}$ indicates a large discrepancy of expression between the two groups. We use $D_{g1,g2}$ and $D_{g2,g1}$ to detect differentially expressed genes.

Summary of the three probe variations for each probe set poses a challenging task because of the enormously large volume of data. The problem becomes worse when interaction effect occurs (i.e. differential expressions are not consistent among a probe set). As discussed in the introduction, alternative splicing and cross-hybridization are two possible factors inducing the interaction effect. Because of the complexity of the interaction effect, we focus on the mild case due to cross-hybridization and alternative splicing. By mild cross-hybridization, we mean that cross-hybridization causes only a small portion of probes with unexpectedly high intensity in a gene. Regarding mild alternative splicing, we mean that alternative splicing causes a few probes to be unable to bind the target gene and results in these probes showing low intensity. The assumption is realistic for mild cross-hybridization because current gene chip techniques have been greatly improved to reduce the cross-hybridization effect. The chance is rare to cause most probes in high intensity levels by cross-hybridization. The assumption is also useful for development of a procedure to identify partially differentially expressed genes due to mild alternative splicing.

With the assumption held, we use a probe level threshold, $P_{probe,E}$, and a gene level threshold, $P_{probe\text{.set},E}$, to select differentially expressed genes and invariant genes. Here $E$ is an indicator of the three measurements of probe variation. Specifically, for selection of invariant genes, we calculate the number of probes with their $D_{g1,g2}$ less than the probe threshold, $P_{probe\text{.set},D_{g1,g2}}$, in a gene. The corresponding percentage, i.e. the calculated number divided by the total number of probes for the gene, is then compared to the gene level threshold, $P_{probe\text{.set},D_{g1,g2}}$. If the percentage is not less than $P_{probe\text{.set},D_{g1,g2}}$, the gene is considered as an (pre) invariant gene. The mathematical formula is expressed by $P(D_{g1,g2} < P_{probe\text{.set},D_{g1,g2}}) \geq P_{probe\text{.set},D_{g1,g2}}$. Genes not in the list of (pre) invariant genes are used to identify differentially expressed genes. For this group of genes, we compute the percentage of probes with $D_{g1,g2}$ greater than $P_{probe\text{.set},D_{g1,g2}}$, and the percentage of probes with $D_{g2,g1}$ greater than $P_{probe\text{.set},D_{g2,g1}}$, for a gene. The two percentage numbers are then compared to $P_{probe\text{.set},D_{g1,g2}}$, and $P_{probe\text{.set},D_{g2,g1}}$, respectively. If one of the two percentages is not less than the corresponding gene level threshold, the gene is considered to have differential expression. The mathematical formulae are given by $P(D_{g1,g2} > P_{probe\text{.set},D_{g1,g2}}) \geq P_{probe\text{.set},D_{g1,g2}}$ and $P(D_{g2,g1} > P_{probe\text{.set},D_{g2,g1}}) \geq P_{probe\text{.set},D_{g2,g1}}$. The process is repeated once with smaller values of thresholds for the pre-invariant genes to choose the final invariant genes and additional differentially expressed genes. The flow chart of the algorithm is given in Figure 1.

**DATA DESCRIPTION**

The prolactin (PRL) gene has been indicated as an important factor on the regulation of the outgrowth of new capillary blood vessels, a process referred to as angiogenesis (Ochoa et al., 2001). In particular, 23 kDa prolactin (23k PRL) has been shown to be angiogenic while its proteolytic fragment 16 kDa prolactin (16k PRL), a fragment of 23k PRL derived from a cleavage at its large disulfide loop, to be antiangiogenic (Kim et al., 2003). We undertook an expression profiling approach to address how signaling by 23k and 16k PRL affects endothelial cell function. We generated recombinant adenovirus expressing the 23k PRL (Ad-23k PRL) and 16k PRL (Ad-16k PRL) and used them to infect human umbilical vein endothelial cells (HUVECs) for 48 h. HUVECs infected with Ad-luciferase (Ad-Luc) were used as a control. Gene expression profile was analyzed by DNA microarray using U95Av2 arrays (Affymetrix) containing 12 600 genes and ESTs. Two arrays from replicates of 16k PRL sample and two from replicates of 23k PRL were compared to two arrays from two replicates of control sample.

This study poses a special property for us to evaluate our approach. That is, 16k and 23k PRLs work as a treatment
Flow Chart of Algorithm

1. Computation of Probe Measurement:

   Calculate PM’s ranks for each array
   Standardize PM’s ranks (denoted by P-rank)
   (i.e. PM rank divided by total number of probes in an array)
   Given P-ranks among all arrays for a probe,
   Compute

<table>
<thead>
<tr>
<th>(Overall variation)</th>
<th>(Between group variation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{range}$</td>
<td>$D_{g1,g2}$</td>
</tr>
<tr>
<td>($A^{th}$ percentile of P-rank among all arrays)</td>
<td>($B^{th}$ percentile of P-rank in the g1 group arrays)</td>
</tr>
<tr>
<td>Where 0≤B≤A≤100</td>
<td>($B^{th}$ percentile of P-rank in the g2 group arrays)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(Pre-invariant Genes)</th>
<th>(Invariant Genes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(D_{range} &lt; P_{probe,D_{range}})$ ≥ $P_{probe,ext,D_{range}}$</td>
<td>$P(D_{g1,g2} &gt; P_{probe,D_{g1,g2}})$ ≥ $P_{probe,ext,D_{g1,g2}}$</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>or</td>
</tr>
<tr>
<td>$P(D_{g1,g2} &gt; P_{probe,D_{g1,g2}})$ ≥ $P_{probe,ext,D_{g1,g2}}$</td>
<td>$P(D_{g2,g1} &gt; P_{probe,D_{g2,g1}})$ ≥ $P_{probe,ext,D_{g2,g1}}$</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

2. Gene Selection:

   $P(D_{range} < P_{probe,D_{range}})$ ≥ $P_{probe,ext,D_{range}}$
   Yes
   (Pre-invariant Genes)

   $P(D_{g1,g2} > P_{probe,D_{g1,g2}})$ ≥ $P_{probe,ext,D_{g1,g2}}$
   or
   $P(D_{g2,g1} > P_{probe,D_{g2,g1}})$ ≥ $P_{probe,ext,D_{g2,g1}}$
   Yes

   (Invariant Genes)

   $P(D_{g1,g2} > P_{probe,D_{g1,g2}})$ ≥ $P_{probe,ext,D_{g1,g2}}$
   or
   $P(D_{g2,g1} > P_{probe,D_{g2,g1}})$ ≥ $P_{probe,ext,D_{g2,g1}}$
   Yes

   Differentially Expressed Genes

**Fig. 1.** Flow chart of algorithm.

factor, but also as a positive control gene. Specifically, 23k PRL is a wild-type human prolactin; its mRNA closely matches to the probe set of the human prolactin gene which is marked as ‘878_s_at’ (abbreviation as prolactin gene thereafter) in Affymetrix U95v2 chip. As a result, the expression of the prolactin gene is completely differential for the 23k PRL. In contrast, 16k PRL is a modified prolactin derived from 23k PRL with some sequences truncated, the similarity of 16k PRL mRNA to the prolactin gene will be weaker. Therefore, the expression level of 16k PRL is lower than 23k PRL (Kim et al., 2003). More importantly, the gene expression becomes partially differential. This precious biological information enables us to investigate two important aspects of gene expression: main treatment effect (complete differential expression) and interaction effect (partial differential expression). A robust statistical approach should be able to account for these effects on the array analysis.

**EVALUATION**

The MAS and Dchip models are used to compare to our approach. Specifically, three MAS measurements are used for analyses: average log difference, average difference, and Tukey Biweight (TB) mean. We apply a non-linear transformation (lowess function) to normalize the three MAS measurements. One-way ANOVA is then employed to select differentially expressed genes with Bonferroni adjustment of p-value (type I error is set to 0.05) for multiple comparisons of thousands of genes (Bickel and Doksum, 1977), and without Bonferroni adjustment of p-value with type I error 0.001. For Dchip, two models are used: PM–MM model and PM only model. The analyses are done based on the default filter criteria.

For our approach, the A and B parameters of percentile-range are set up as 100 and 0, respectively, because of small sample size. The probe level thresholds of $D_{range}$,
RESULTS

The expression levels in the human prolactin probe set were differential between the two treatment groups (23k and 16k PRLs) and the control groups. The result implied the microarray experiment was reliable at least for the control gene. Specifically, Figure 2(a) shows the majority of probes in this probe set had expressions higher in the 23k PRL than in the control. In contrast, only about half of the probes displayed differential expressions between the 16k PRL and the control in Figure 2(b). The expression patterns matched the biological background of the 23k PRL as a wild type of human prolactin and the 16k PRL as a modified 23k PRL with C-terminal-sequences truncated.

Table 1 is the comparison results of identification of the control gene and of the number of selected genes. Our approach identified the prolactin probe set as a differential expressed gene between the two treatment groups (23k and 16k PRLs) and the control groups. In contrast, Dchip and MAS (using one-way ANOVA without Bonferroni adjustment of p-value) were able to pick up the probe set in the comparison of the 23k PRL and the control, but failed to identify the probe set in the comparison of the 16k PRL and the control. MAS using one-way ANOVA with Bonferroni adjustment of p-value did not find any differentially expressed genes.

In terms of number of selected genes, our approach identified 63 and 60 differentially expressed genes for the comparisons of the 23k and 16k PRLs versus the control, respectively. Dchip and MAS had smaller numbers of selected genes for both comparisons (19–46 for the 23k PRL case and 8–24 for the 16k PRL case). Only a small portion of their selected genes were in the list of our selected genes.

Table 1. Comparison of different statistical approaches for identification of the positive control gene

<table>
<thead>
<tr>
<th>Method</th>
<th>Measurement</th>
<th>Status of identification</th>
<th>23k PRL versus control</th>
<th>16k PRL versus control</th>
<th>Overlap&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our approach</td>
<td>Rank of PM</td>
<td>Yes (63)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Yes (60)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dchip PM–MM</td>
<td>PM–MM</td>
<td>Yes + (28, 8)</td>
<td>No (14, 2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dchip PM</td>
<td>PM</td>
<td>Yes + (28, 8)</td>
<td>No (14, 2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One-way ANOVA without Bonferroni adjustment (type I error = 0.001)</td>
<td>MAS average log difference</td>
<td>No (21, 0)</td>
<td>No (12, 0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One-way ANOVA with Bonferroni adjustment (type I error = 0.05)</td>
<td>MAS average log difference</td>
<td>Yes (19, 2)</td>
<td>No (24, 2)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Yes: identify the positive gene (human prolactin gene). No: fail to identify the positive control gene.

Table 2. Comparison of different statistical approaches for PCR confirmatory test of five genes

<table>
<thead>
<tr>
<th>Method</th>
<th>Measurement</th>
<th>23k PRL versus control</th>
<th>16k PRL versus control</th>
<th>Overlap&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our approach</td>
<td>Rank of PM</td>
<td>4&lt;sup&gt;b&lt;/sup&gt; (5&lt;sup&gt;b&lt;/sup&gt;)</td>
<td>3 (3)</td>
<td>3</td>
</tr>
<tr>
<td>Dchip PM–MM</td>
<td>PM–MM</td>
<td>3 (4)</td>
<td>1 (1)</td>
<td>1</td>
</tr>
<tr>
<td>Dchip PM</td>
<td>PM</td>
<td>2 (3)</td>
<td>1 (1)</td>
<td>1</td>
</tr>
<tr>
<td>One-way ANOVA without Bonferroni adjustment (type I error = 0.001)</td>
<td>MAS average log difference</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Overlap represents the number of confirmed genes in both 16 k and 23 k PRLs.

<sup>b</sup>The number of confirmed genes.

Five genes based on our selection were chosen for PCR tests. The results were given in Table 2. Four of the five genes were confirmed as differentially expressed genes in both cases of 16k and 23k PRLs. Only one gene showed false positive. Our approach had three confirmed genes in the list of 16k PRL and four confirmed genes in the list of 23k PRL. In contrast, Dchip had a smaller number of confirmed genes (1 in 16k PRL and 2–3 genes in 23k PRL). Our approach and Dchip had the
same false positive gene. All the five genes were in not in the list using the MAS measurements.

Results of selection for the invariant genes showed that 535 and 388 genes were identified as the invariant genes for the comparisons of the 23k and 16k PRLs versus the control, respectively. There were 321 genes (83%) selected in both comparisons. Affymetrix housekeeping genes (Affymetrix id: HSAC07/X00351_3, HSAC07/X00351_M,
Table 3. Comparison of linear normalization coefficients between using the invariant set of genes versus the whole set of genes

<table>
<thead>
<tr>
<th></th>
<th>Control 1</th>
<th>Control 2</th>
<th>23k 1 PRL</th>
<th>23k 2 PRL</th>
<th>16k 1 PRL</th>
<th>16k 2 PRL</th>
<th>23k PRL versus control</th>
<th>16k PRL versus control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>S</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Invariant genes (535)</td>
<td>0.05</td>
<td>0.98</td>
<td>0.12</td>
<td>1.00</td>
<td>0.80</td>
<td>0.96</td>
<td>0.45</td>
<td>0.98</td>
</tr>
<tr>
<td>All genes (12,625)</td>
<td>−0.02</td>
<td>0.99</td>
<td>0.24</td>
<td>0.98</td>
<td>0.03</td>
<td>1.04</td>
<td>−0.15</td>
<td>1.04</td>
</tr>
<tr>
<td>Control 1</td>
<td>I</td>
<td>S</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Invariant genes (388)</td>
<td>0.02</td>
<td>0.98</td>
<td>0.19</td>
<td>1.00</td>
<td>0.35</td>
<td>0.97</td>
<td>0.26</td>
<td>0.93</td>
</tr>
<tr>
<td>All genes (12,625)</td>
<td>−0.02</td>
<td>0.99</td>
<td>0.24</td>
<td>0.98</td>
<td>−0.39</td>
<td>1.04</td>
<td>0.83</td>
<td>0.87</td>
</tr>
</tbody>
</table>

I. Intercept; S, Slope.

HUMGAPDH/M33197_3, HUMGAPDH/M33197_5 and some spike control genes (Affymetrix id: BioDn-3, CreX-3, CreX-5) were in the list. Table 3 is the list of linear normalization coefficients using the invariant set of genes and the whole set of genes. All intercepts were close to 0 and all slopes were near to 1.

DISCUSSION

To model gene expression, data are often generated by using spike-in genes with known concentration inserted in the sample. However, the approach is questionable because the spike-in genes may not reflect the true information of research interest in the system (Schadt et al., 2002). In contrast, the data set from this study presents a unique feature to address the issue. The study used the prolactin gene as a treatment factor in the 16k and 23k PRLs samples with a different expression form. Because the gene chip contains the gene, the prolactin gene naturally becomes a control gene. Therefore, we are able to assess the reliability of the experiment. Most importantly, the property allows us to evaluate this gene expression either in the case of the main effect or interaction effect. In addition, the PCR tests provide confirmatory results to examine the prediction performance. Though the study had a small sample size, but it is a typical setting of microarray experiments in basic medical research. Therefore, the data possess various valuable information in evaluating our approach.

Results have shown our approach outperformed the other approaches either in the selection of the control gene or in the prediction performance. The explanations are given as follows.

The profile of the prolactin gene in Figure 2(a) displays all probe intensities except the probe 16 are substantially higher (at least 4-fold change) in the 23k PRL than in the control. Differences of expressions are consistent among the 15 probes. The observation implies interaction between the probe effects and the treatment effect is mild compared to the main treatment effect: in other words, magnitude of probes binding to the target gene is most contributed by the treatment effect so that differences of expressions are homogeneous in the same series of probes. The phenomenon fits the assumption of the additive model by Dchip and MAS. Therefore, both approaches are able to distinguish differences of the prolactin expressions between the 23k PRL and the control.

However, when expression alternations depend on the probe effects, these two approaches without accounting for the interaction effect is likely to fail. The 16k PRL is the case that demonstrates a situation of interaction between the probe affinity and the treatment effect as seen in Figure 2(b). Only about half of the probes of the human prolactin probe set have higher expressions in the 16k PRL than in the control arrays. The other probes in the same series of probes show low expressions in both groups. The heterogeneity of difference of probe expressions causes the MAS and Dchip models unable to detect expression changes in this control gene. With the assumption of homogeneity of expression changes in the same series of probes, both approaches treat all probes in the set receiving the same amount of treatment effect. As a result, the overall expression is a compromised expression between probes with high expressions and probes with low expressions, and therefore, becomes biased.

For our approach, we use $P$-rank as the primary outcome variable for analysis. The use of the measurement does not only work as a standardized score for array comparison, but is also used to reduce the effect of extremely high intensities. The percentile-range is introduced to analyze $P$-rank in order to measure overall variation and between group variation. By using two flexible thresholds to account for outliers, the percentile-range provides a useful tool to differentiate probe expressions. With employment of the two filters for gene selection, we compute the percentage of probe expression changes to examine similarity of mRNA between the target gene and the gene chip’s probe set. If there is no interaction, the degree of similarity of mRNA tends to be high. As a result, the percentage of probe expression changes is close to 100%. In contrast, when interaction occurs, some probes may not respond to the target gene. Consequently, the magnitude of similarity decreases and the percentage becomes smaller. As a result, our integrated approach is able to identify the prolactin gene as a differential expressed gene either in the case of only main treatment effect (i.e. complete differential expressions) or in the case of interaction effect (i.e. partial differential expressions).

In addition, the PCR confirmatory test has shown that our approach has better performance than the other approaches in terms of prediction of regulated genes. Among the five selected genes, only one false positive gene was found in the case of 23k PRL and one false negative gene was in the case of 16k PRL by using our approach. In contrast, Dchip
had a larger number of false negative genes in both cases (1–2 and 3 for 23k and 16k PRLs, respectively). Particularly, Dchip had only one confirmed gene in its list for the case of 16k PRL. As discussed before, the prolactin gene was a treatment factor in the study. Its expression is partially differential in the case of 16k PRL. This interaction effect might cause Dchip to be less efficient in identifying genes with expression changes.

Analysis of the invariant sets of genes suggests that our procedure is able to adequately select invariant genes and provides an effective way for normalization. Specifically, the procedure identifies housekeeping genes and some spike control genes as the invariant genes in the comparisons of the 23k and 16k PRLs versus the control. In addition, the degree of consistency of the invariant genes remains high between the two comparisons (83% of the invariant genes are overlapped). Moreover, the selected invariant genes are only 3–4% of the whole genes, but their linear coefficients for normalization are comparable to those coefficients generated by the use of the whole genes as seen in Table 3. The performance implies our procedure is efficient for normalization because only a small percent of genes are used. Furthermore, in contrast to the whole genes, the invariant genes are unlikely to include differentially expressed genes, so the regulated genes have less impact on the normalization curve.

In summary, we have developed a probe-based approach for the oligonucleotide array data analysis. The approach integrates rank, percentile-range, and filter of expression changes. The analyses from the demonstrated data have shown our approach as a useful tool to identify gene expression alternations even in the case of alternative splicing.

ACKNOWLEDGEMENTS

This work was supported in part from the National Cancer Institute (5P30 CA-13148 and 1P50 CA89019). The authors gratefully acknowledge the constructive criticisms by the anonymous referees.

REFERENCES


