Gene expression

Differential and trajectory methods for time course gene expression data

Yulan Liang¹,*, Bamidele Tayo², Xueya Cai¹ and Arpad Kelemen¹,³
¹Department of Biostatistics and ²Department of Social and Preventive Medicine, University at Buffalo, Buffalo, NY 14226, USA and ³Department of Computer and Information Sciences Niagara University, Lewiston, NY 14109, USA

Received on December 13, 2004; revised on April 22, 2005; accepted on April 23, 2005
Advance Access publication May 10, 2005

ABSTRACT

Motivation: The issue of high dimensionality in microarray data has been, and remains, a hot topic in statistical and computational analysis. Efficient gene filtering and differentiation approaches can reduce the dimensions of data, help to remove redundant genes and noises, and highlight the most relevant genes that are major players in the development of certain diseases or the effect of drug treatment. The purpose of this study is to investigate the efficiency of parametric (including Bayesian and non-Bayesian, linear and non-linear), non-parametric and semi-parametric gene filtering methods through the application of time course microarray data from multiple sclerosis patients being treated with interferon-β-1a. The analysis of variance with bootstrapping (parametric), class dispersion (semi-parametric) and Pareto (non-parametric) with permutation methods are presented and compared for filtering and finding differentially expressed genes. The Bayesian linear correlated model, the Bayesian non-linear model and the non-Bayesian mixed effects model with bootstrap were also developed to characterize the differential expression patterns. Furthermore, trajectory-clustering approaches were developed in order to investigate the dynamic patterns and inter-dependency of drug treatment effects on gene expression.

Results: Results show that the presented methods performed significantly but all were adequate in capturing a small number of the potentially relevant genes to the disease. The parametric method, such as the mixed model and two Bayesian approaches proved to be more conservative. This may because these methods are based on overall variation in expression across all time points. The semi-parametric (class dispersion) and non-parametric (Pareto) methods were appropriate in capturing variation in expression from time point to time point, thereby making them more suitable for investigating significant monotonic changes and trajectories of changes in gene expressions in time course microarray data. Also, the non-linear Bayesian model proved to be less conservative than linear Bayesian correlated growth models to filter out the redundant genes, although the linear model showed better fit than non-linear model (smaller DIC).

We also report the trajectories of significant genes—since we have been able to isolate trajectories of genes whose regulations appear to be inter-dependent.

Availability: SAS, R and WinBugs codes are available upon request from the authors.
Contact: yliang@buffalo.edu

1 INTRODUCTION

Satisfying the computational demand for statistical analysis of gene expression data has been a great challenge, given the usual large number of genes or probes that are involved. This problem becomes more daunting due to the multi-collinearity among genes, or due to the high correlations among the experimental conditions in the temporal fashions. Also, the type of statistical analysis required increases in complexity such as in the mixed model where analysis of variance (ANOVA) involves several main factors and many interaction terms. To mitigate these problems, some gene filtering and wrapping methods have been proposed by eliminating the redundant genes based on set criteria prior to or during the statistical differential gene expression analysis. Tseng et al. (2001) proposed using a quality index computed from duplicate spots on the same slide, to filter out outlying spots and poor quality genes. A direct significant analysis to select genes based on t-test or regularized t-statistics were proposed by Tusher et al. (2001) and Baldi and Long (2001). Keller et al. (2000) used likelihood-based metrics to select the genes for a naive Bayes classifier for classifying tissues. Yeung and Ruzzo (2001) used principal component analysis to extract gene components sequentially, which can maximize the total predictor variability. Partial least square (orthogonal) components proposed by Ghosh (2003) were constructed to maximize the sample covariance between the response and the linear combination of the gene expression values. Xiao et al. (2004) proposed multivariate test statistic and algorithm for identifying differentially expressed gene combinations. Genetic and evolution based gene selection schemes were proposed by Ooi and Tan (2003). Pochet et al. (2004) and Romualdi et al. (2003) applied wrapper methods with various variable selection methods such as linear and non-linear dimension reductions with artificial neural networks for classifications. Head to head comparisons of differential gene expression have also been conducted: Pan (2002) compared three parametric methods: the t-test, regression modeling and a mixture model approach to discover differentially expressed genes in replicated microarray experiments. Troyanskaya et al. (2002) compared three non-parametric methods for the same purpose.

While some of these methods have been proved to be useful in the analysis of single-point gene expression data on distinguishing different cell types, not much has been achieved in the application and comparison of various gene filtering methods or differential methods for time course gene expression data. Furthermore, cells that appear similar may follow different clinical behavior, such as
the response to treatment drugs under different experimental conditions or time points and so can be dramatically different. Therefore, another important issue in microarray analysis, which has received less attention, is the differentiation of time course gene expression, with sequential time points, under different experimental conditions. This becomes an important source of information for drug development and efficient drug therapies.

Although the coordinated (temporal) gene expression study was mainly preoccupied with unsupervised clustering, efficient gene filtering and differentiation approaches, based on time course gene expression prior to clustering analysis, can help to remove redundant genes and noise while highlighting the most relevant genes that are major players in the development of certain diseases. Biological variations of gene expressions and multiple measurements of each gene across time, make it possible to statistically assess the variability of genes. Singular value decomposition proposed by Alter et al. (2000) was developed for the identification of characteristic modes that represent temporal changes of gene expression. Hastie et al. (2000) proposed gene shaving with GAP statistic for identifying distinct sets of genes with similar expression patterns.

In this paper, we study the efficiency of various parametric, non-parametric and semi-parametric methods in Bayesian and non-Bayesian settings and linear and non-linear models for filtering, identification of significant genes and characterizing the dynamics of gene expression in time course gene array data. The word efficiency is used in this study to mean both the quantity, such as the effectiveness of reducing the number of genes (parameters) in the models, while controlling the type I error and the quality of the filtering methods which indicate the reproducibility of the methods. We present and compare six different gene filtering and differentiation methods for time course gene expression data. These methods include parametric (ANOVA with bootstrapping, mixed model, linear and non-linear Bayesian models), semi-parametric (class dispersion) and non-parametric (Pareto with permutation) methods. The genes discovered by the different methods were further verified with the genes known to be associated with the disease to check the reproducibility of the methods. In addition, Deviance Information Criteria (DIC) was used to compare the linear and non-linear Bayesian models’ performances (Spiegelhalter et al., 2002).

Moreover, trajectory-clustering approaches were developed for characterizing the dynamic patterns and inter-dependency of drug treatment effects on gene expression. The genomic effects of interferon (IFN)-β-1a treatment on lymphocytes in the peripheral blood of multiple sclerosis (MS) patients (14) using 10 time-point microarray data and 4324 probe sets prepared by Weinstock-Guttman et al. (2003) were evaluated and analyzed through our approaches.

2 METHODS

2.1 Analysis of variance with bootstrap method

We fitted probe-specific one-way ANOVA models using the general linear model procedure in SAS with time as the main factor. The main interest was to investigate and select probes with high variability in expression between time points. We then carried out a bootstrap resampling with 1000 replicates on the p-values of the F-statistics from ANOVA.

Let \( p = (p_1, p_2, \ldots, p_n) \) denote the original sample of p-values of the F-statistics from the ANOVA models. We created B bootstrap samples \( \hat{p}^1, \ldots, \hat{p}^B \) by sampling with replacement, each of size \( n \) equal to 4324 and B replicates equal to 1000. For each bootstrap sample \( \hat{p}^b = (\hat{p}^b_1, \hat{p}^b_2, \ldots, \hat{p}^b_n), b = 1, 2, \ldots, B \), we computed \( \hat{p}_i \) for the \( i \)-th gene.

\[
\hat{p}_i = \frac{1}{B} \sum_{b=1}^{B} \hat{p}^b_i, \quad i = 1, \ldots, n.
\]

\[
s^2 = \left[ \sum_{b=1}^{B} (\hat{p}^b_i - \hat{p}_i)^2 / (B-1) \right]^{1/2}.
\]

Using \( \alpha = 0.025 \), we obtained an estimate of the \( 1 - \alpha \) percentile from empirical distribution of bootstrap sample of p-values. This bootstrapped estimate was then used as cut-off point for selecting probes with \( p \leq 0.000005 \) (the estimate) while those with greater \( p \)-values were filtered out.

2.2 Two-interconnected mixed models with Bootstrap methods

Two-interconnected mixed models: the ‘normalization’ model and the ‘gene’ model proposed by Wolfinger et al. (2001) for each gene at each time point were conducted in our analysis. Both models are similar to the overall ANOVA model of Kerr et al. (2000), in that they model the logarithms of the original fluorescence measurements not log-ratio values. The normalization model accounts for experiment-wide systematic effects that could bias inferences made on the data from the individual genes: \( \log(Y) = \mu + T_k + P_i + (PT)_{ki} + \varepsilon \). \( \log(B) \) is the global normalized gene expression data for patient \( i \), gene \( j \) at time point \( k \), \( \mu \) represents an overall effect, \( T \) is the main effect of time points, \( P \) is the main effect of patients, \( \varepsilon \) is stochastic error.

In this model, the main effect of patients, the interaction effect of time points and patients and the stochastic error are considered as random effects, and the remaining parts are considered as fixed effects. The residuals from this model represent normalized values and are the input data for the gene-specific model.

The gene-specific models were fitted separately to the normalized data from each gene, allowing inferences to be made, using separate estimates of the variability: \( r_{jk} = G_j + (GT)_{jk} \). \( r_{jk} \) are the residuals from the previous normalization model, computed by subtracting the fitted values for the effects from the \( Y_{ijk} \) values. Note that in this model, the effects are analyzed gene by gene. \( PG \) is the effect for each gene for each patient. The \( GT \) effects are measured for each gene at each time point, which is of our primary interest. We tested the differences between these effects based on t-tests of all possible pairwise comparisons within a gene. \( \gamma_{jk} \) is the stochastic error. The restricted maximum-likelihood (REML) method was used to estimate the variance components that also produced estimates of all time points in the overall ANOVA model of Kerr et al. (2001) for each gene at each time point.

\[
Y_{ijk} \sim \text{Normal}(\mu_{ijk}, \tau_{ijk}), \quad (1)
\]

\[
\mu_{ijk} = \beta_1 \cdot X + \beta_{2j} \cdot X \sim \text{MVN}(\mu_{ijk}, \Omega), \quad (2)
\]

\( Y_{ijk} \) is the measurement from patient \( i \), gene \( j \) at time \( k \). \( X \) stands for time points. Genes are considered as random effects. \( \beta_1 \) and \( \beta_{2j} \) are gene-specific random effects with bivariate Gaussian distribution. This allows us to test the possible dependency of the temporal process of expression on the pre-treatment expression level. For example, higher initial expression may lead to higher subsequent expression (positive correlation) or lower subsequent expression (negative correlation). Similarly, we assume non-informative independent univariate normal priors [Normal (0, \( 10^{-6} \))] for the components \( \mu_{\beta_1} \) and \( \mu_{\beta_{2j}} \). A Wishart (\( K, \rho \)) prior is specified for \( \Omega \), the population covariance matrix.
precision matrix of the regression coefficients. To represent vague prior knowledge, we chose the degree of freedom for this distribution to be 2, the same as the rank of $\Omega$. The scale matrix was specified as

$$R = \begin{pmatrix} 100 & 0 \\ 0 & 0.1 \end{pmatrix},$$

which demonstrates the prior guess at the order of magnitude of the covariance matrix $\Omega^{-1}$ for $\beta_j$.

### 2.4 Non-linear Bayesian time effects model

A higher level of sophistication in Bayesian modeling of the gene expression is to relax the linear assumption on the time effects and extend (2) in the linear Bayesian growth model to non-linear setting. This is achieved by modeling separate random gene effects as piece-wise expression changes at each time point. The model thus can be expressed as

$$y_{jk} = \beta_{1j} + \beta_{2jk} + \epsilon_{jk},$$

where $y_{jk}$ is the measurement from patient $i$, gene $j$ at time point $k$. $\beta_{1j}$ is the baseline expression for each gene $j$; $\beta_{2jk}$ are the expressions of gene $j$ at time point $k$.

Since genes are more likely to be non-linearly expressed, model (3) may be more appropriate than (2), although its computational intensiveness is inevitable. We assume normal distribution for $\beta_{1j}$ and independent normal distributions for $\beta_{2jk}$, with typical prior assumptions. The distributional assumption for $\beta_{2jk}$ allows unequal variations in gene expressions for each gene at each time point.

Both the above linear and non-linear Bayesian models have the advantage of dealing with irregular time intervals, compared with the dynamic Bayesian clustering method (Ramoni et al., 2002), which used autoregressive models, where the time points have to be evenly spaced. Moreover, both models take into account the multi-level heterogeneous variations of the expression values and the high noise levels. However, the non-linear Bayesian model focuses on the non-linear effect estimates, while the linear Bayesian multivariate growth model addresses the correlation covariance information of time for the identification of genes that respond differently to the IFN-$\beta$-1a treatment and for discovering the temporal patterns changing from time to time.

Since many parameters with multiple levels are involved in the estimation, marginalization such as using marginal posterior mode is more efficient. Monte Carlo Markov chain (MCMC) algorithm, with Gibbs sampling, was implemented to undertake this task of simultaneous sampling of the conditional posterior distributions of all parameters (Congdon, 2001) and were implemented in WinBugs (Bayesian inference using Gibbs sampling). A total of 2000 samples, after 1000 burn-ins, were used for the computation of parameters. In addition, multiple testing problems were carefully handled through regularized $t$-posterior statistics (Baldi and Long, 2001).

### 2.5 Class dispersion: a semi-parametric method

This method is based on both measures, of intra-class dispersion and inter-class dispersion, as described by Fleury et al. (2002). Here, the interest was on probes with significant non-constant trajectories, based on small variability at each time point among patients or replicates; and large variability between populations at different time points.

Let $y_j(t)$ be the expression on the $j$-th gene, $j \in \{1, \ldots, J\}$ in the $i$-th patient, $i \in \{1, \ldots, I\}$ at the $t$-th time point, $t \in \{1, \ldots, T\}$. For our data, $I, J$ and $K$ are 14, 2434 and 10, respectively. The intra-class dispersion of the $j$-th gene at the $t$-th time is based on small variability at each time point among patients, and can be defined as

$$\xi_j^1(t) = \sum_{i \in \mathcal{I}} |y_j^i(t) - \bar{y}_j(t)|,$$

The inter-class dispersion for the $j$-th gene between times $t_1$ and $t_2$ is based on the large variability between populations at each time point.

$$\xi_j^2(t_1, t_2) = \sum_{j \in \mathcal{J}} ||y_j^i(t_1) - y_j^i(t_2)||,$$

where $||\cdot||$ denotes a norm, e.g. $l_1, l_2$ or $l_{\infty}$, and in this case is $l_1$-norm.

To separate any given two time points for the $j$-th gene, we used a test statistic based on the ratio of the two dispersion measures given above and defined as:

$$r_p(t_1, t_2) = \frac{Q - 1}{2Q} \frac{\xi_j^1(t_1) + \xi_j^1(t_2)}{\xi_j^2(t_1, t_2)}.$$

$\tau_p(t_1, t_2) > \tau^*(1 - \alpha)$, where $\tau^*(1 - \alpha)$ is threshold chosen from $1 - \alpha$ percentile of the distribution of permutation resampling of $\tau_p$. This test, which is similar in function to the paired $t$-test, was then used to select gene probes with significant $\tau_p$.

### 2.6 Pareto: non-parametric filtering method

Instead of using a scalar criterion as in Equations (4–6), we modified a non-parametric filtering method, Pareto, proposed by Fleury et al. (2002) to account for the temporal nature of the present data. Specifically, the method was based on joint-maximization of three criteria: monotonicity $\xi_j^1$, end-to-end increase $\xi_j^2$ and mean slope difference $\xi_j^3$ of the gene probe trajectories.

These multiple gene filtering criteria were defined as:

$$\xi_j^1 = \frac{1}{T} \sum_{t \in \mathcal{T}} |y_j^i(t + 1) - y_j^i(t)|,$$

$$\xi_j^2 = \frac{1}{T} \sum_{t \in \mathcal{T}} (y_j^i(t^i) - y_j^i(t^-))(y_j^i(t^i) - y_j^i(t^-)),$$

$$\xi_j^3 = \max_{t \in \mathcal{T}} \left( \left| y_j^i(t + 1) - y_j^i(t^-) \right| - \min_{t \in \mathcal{T}} \left( \left| y_j^i(t + 1) - y_j^i(t^-) \right| \right) \right),$$

where $T$ is the number of time points, $I$ is the number of patients and $T^i$ is the total possible trajectories constructed from different patients, at each time point. For instance, given our data with 14 patients and 10 time points, the total number of time trajectories for each gene is $10^{14}$. Let $\mathcal{S}$ denote the average of $y^i$ over $i$. More discussions on these criteria can be found in Fleury et al. (2002) and Steuer (1986). Genes were selected from the 95th percentile of the ranked distribution of each criterion; and, genes not selected based on all the three criteria were filtered out.

### 3 DATA ANALYSIS

#### 3.1 Data

Multiple sclerosis is an autoimmune disease in which the body’s immune system attacks the brain and spinal cord, resulting in neurological disabilities (Steinman and Zamvil, 2003). This inflammatory-demyelinating disease of the central nervous system affects over 1 million people worldwide. It is a complex, variable disease that causes physical and cognitive disabilities. Recombinant human IFN-$\beta$-1a is one of the most commonly prescribed forms of therapy for relapsing MS patients on the basis of several clinical trials (Jacobs et al., 1996; Weinstock-Guttman et al., 2003). The effects of IFN-$\beta$-1a treatment are complex, and its pharmaco-dynamics at the genomic level in humans are poorly understood.

The study population, cell analysis, RNA isolation and DNA arrays were described by Weinstock-Guttman et al. (2003). In this study, 14 patients with active relapsing remitting MS were recruited. These patients had not previously received IFN-$\beta$-1a and were clinically stable for the preceding 4 weeks. Peripheral blood samples were obtained from all patients just before treatment and at 1 h, 2 h, 4 h, 8 h, 24 h, 48 h, 5 days, 7 days and 3 months after the administration of a 30 $\mu$g dose of intra-muscular IFN-$\beta$-1a, and processed for subsequent gene expression array using GeneFilters GF211 DNA arrays (Research Genetics, Huntsville, AL). The final normalized data contained 4324 probes per time point and per patient.
Given the large number of gene probes, the major task was to reduce the number by filtering out redundant probes, based on their expression data. Weinstock-Guttman et al. (2003), using candidate gene approaches, specifically examined the lymphocyte gene expression response in five antiviral genes (protein kinase, Mx2, Mx1, GBP-1 and GBP-2), six IFN-β-1a signaling genes (IFNAR1, IFNAR2, Jak1, Tyk2, Stat1 and p48) and five lymphocyte activation genes (1-8D, 1-8U, β2-microglobulin, CD69 and 9-27). As a follow up study to the experiment, we implemented our proposed methods: the ANOVA with bootstrap, mixed models, linear and non-linear Bayesian growth models, class dispersion with permutation and Pareto methods, independently for the whole genome wide dataset.

3.2 Results

3.2.1 Filtering methods For the two-interconnected mixed models, since there were 10 time points, totally there were 45 pairwise t-tests within each gene with degree of freedom of 116. The p-value cut-off point was $0.05/(4324 \times 45) = 2.6 \times 10^{-7}$, and 54 genes were found significantly expressed at certain time points. However, with the criteria of 4-fold change, only a few genes were found and most of them were false positives, since a large fold change by itself does not consider the variance. The significantly expressed genes, which were found by the two-interconnected mixed models were ignored by using the fold change criteria. This result showed the failure of the fold change criteria, and also showed that the gene-by-gene variance estimates provided statistically appropriate backdrops to assess the significance of the difference within time points. An interesting result was that most of the significant changes happened from pre-treatment to 2 h after treatment, and from 8 to 48 h after treatment (Fig. 1). The selection result from semi-parametric method: class dispersions with permutation are shown in Figure 2, with the positions of selected genes shown in red at each time point.

![Fig. 1. Significant changes at different time points. y-axis: number of significant genes. x-axis: time changes—current versus previous time point.](https://academic.oup.com/bioinformatics/article-abstract/21/13/3009/197421)

![Fig. 2. Plots of genes’ inter-class dispersion against intra-class dispersion for the nine time points after drug treatment. Positions of selected genes based on the class dispersion (semi-parametric) method are shown in red at each time point.](https://academic.oup.com/bioinformatics/article-abstract/21/13/3009/197421)
When applying linear Bayesian model and the regularized $t$-posterior test to control the Type I error (0.05 overall), the adjusted $\alpha$-level was $0.05/4324 = 1.156 \times 10^{-5}$. The critical value of the $t$-statistic was thus 4.5635, with a degree of freedom of 126 ($14 \times 9$: 14 patients and 9 time points after treatment). 435 genes were filtered as significantly expressed at least once, and 340 all the time. For the non-linear Bayesian model based on Equation (3), the critical value of the $t$-statistic for the regularized two-sample $t$-test was 6.2585, with adjusted $\alpha$-level $0.05/(4324 \times 9) = 1.28 \times 10^{-6}$. Using this model, 126 genes were selected, for which at least one differential expression at each of the time points after treatment was detected; among these genes, 26 genes had differential expressions over the entire time span after treatment.

For comparing the efficiency of various methods, we first compared the number of genes selected by each of the six methods and looked at the common genes among them. The highest number of genes was selected by the linear multivariate Bayesian model (435 at least one time sig.; 340 all time sig.), followed by non-linear Bayesian (126), class dispersion (123), ANOVA (108), Pareto (88) and mixed model (54). The ANOVA and class dispersion methods had 34 genes in common, while Pareto filtering had just one gene in common with the class dispersion method (Fig. 3). The lack of genes commonly identified by these three methods could be attributed to two methodological differences, the stringency from multiple criteria such as from Pareto and the poorly understood effects of IFN-$\beta$ (Rudick et al., 1998; Gayo et al., 1999).

Second, in order to further compare the discussed methods, we varied the $\alpha$ significant level (Type I error) from 0 to 1 and plotted the proportion of rejections of the null hypothesis (from 0 to 1) against the $\alpha$-level for each method. Here we used proportion of rejections of the null hypothesis, instead of the number of rejected hypothesis as in Dudoit et al. (2003), due to the difference of the total number of hypotheses among the six methods. However the two-interconnected mixed models (MM), the Bayesian linear model (BMN) and the Bayesian Non-linear model (BNM) behaved more conservatively than the other three (ANOVA, class dispersion and Pareto), in which the simulation curves became flat very quickly, when $\alpha$ started to increase; therefore, we plotted these three methods (Fig. 4): two-interconnected mixed models (MM), Bayesian linear multivariate model (BMN) in one figure (lower panel); and ANOVA, class dispersion and Pareto in the other figure (upper panel). In the upper panel, we retained the scales and values of $x$- and $y$-axes from 0 to 1, while in the bottom subfigure we reduced the $y$-axis scale from 1 to 0.125, for better visualization in the difference in the methods.

Results show that the class dispersion method is the most robust and performs best in comparison to the other methods with regard to the given different Type I error rates, the proportion of rejection of the hypothesis. ANOVA with bootstrap and Pareto behaved very similarly and performed relatively well. The Bayesian linear multivariate model (BMN) and the Bayesian Nonlinear Model (BNM) behaved more conservatively than the previous three methods. Two-interconnected mixed models (MM) performed most conservatively. This may be due to the fact that two-stage mixed effect models may tease out more genes, including the experiment-wise artifacts and noises. The DIC values for linear Bayesian model (DIC = 1 476 330) compared with non-linear Bayesian model (DIC = 1 509 140) showed the better fit of the linear Bayesian model for the given data.
Fig. 5. Significant monotonic changes in expression (left-hand side) and trajectories of expression (right-hand side) of four representative genes selected by the Pareto method.
and class dispersion methods, simultaneously. Four additional genes (Hs.183487, Hs.241510, Hs.21486 and Hs.20315) were found by our ANOVA method and one additional (Hs.75516) by our class dispersion method. Only two of these genes were identified as differentially expressed in the mixed model (Hs.146360 and Hs.274382). Variations in the identified genes could be a consequence of the different model assumptions and algorithms of the methods.

3.2.2 Post-filtering analysis: trajectory analysis To investigate differentially regulated genes selected by the above methods, we sorted out significantly up- and down-regulated genes, at each time point. These results were then used to display the trajectories of selected genes with significant monotonic changes across specified time points. The selected genes with significant monotonic changes in their expression across multiple time points were further studied through trajectory analysis. Some representative genes and their patterns of changes are plotted in Figure 5. These genes were categorized into three large groups: up-regulate at the first 4 h, then retain the same expression level; up-regulate in the first 4 h, then down-regulate; down-regulate in the first 4 h, then retain the same expression level. In addition, gene names Hs.4217 and Hs.1369 were observed to have a rather consistent up-regulation and down-regulation, respectively, across the entire time series. Two genes (Hs.176663 and Gene H20822 (Fc fragment of IgG, low-affinity IIIa, receptor for (CD16)) maintained almost the same expression level during the treatment; and some other genes (Hs.10306 and Hs.105806, T57859 (natural killer cell group 7 sequence), and AA70165 (granulysin)) up-regulated in the first 4 h, then maintained the same expression level for a period, and then gradually up-regulated again.

Furthermore, we carried out a clustering of the selected genes into three different groups of inter-dependence, based on the onset time of regulation after treatment. Similarly, a possible inter-gene dependence pathway was proposed, based on semi-parametric (class dispersion) methods. A total of 99 genes were found to have onset time of regulation within the first 8 h after treatment, and hence were labeled ‘early’ group; those with onset around 24 h after treatment were labeled ‘intermediate’ group; and those with onset time of or beyond 48 h were labeled ‘late’ group (Fig. 6). Identified genes found in the ‘early’ group include Hs.62661, Hs.274382, Hs.146360 and Hs.183487. Regulation of genes in the ‘late’ group was sustained for longer durations ranging from 2 days to 3 months. These delayed inducements and the prolonged expressions of genes are thought to be some of the pharmacodynamics of IFN-β which are still poorly understood, but on which more data are accumulating.

4 DISCUSSION

In this study, we have developed and compared parametric, semi-parametric and non-parametric gene filtering methods to reduce the number of genes in a temporal microarray data. Also, resampling techniques, including permutation and bootstrap techniques, were incorporated in order to improve statistical inference. The different methods of gene filtering—parametric, semi-parametric and non-parametric—in Bayesian and non-Bayesian settings and with linear and non-linear models performed significant differently, but all were adequate in capturing a small number (<10%) of the potentially relevant genes from a total of 4324 genes.

In order to further examine the various methods, we computed the proportion of rejections of the hypothesis against the α-level for each method by varying the significant α-level from 0 to 1; and the results also showed some significant differences among the different methods. The parametric method, such as the mixed models and the two Bayesian approaches proved to be more conservative. This may due to these methods being based on overall variation in expressions across all time points. The semi-parametric (class dispersion) and non-parametric (Pareto) methods are appropriate for capturing variation in expressions from time point to time point, thereby making them more suitable for time course gene array data. Although, the modified Pareto method has more stringent combined multiple criteria it still less conservative, and filtered more significantly expressed genes than the mixed effects models and the two Bayesian methods. Each method appears to be differentially sensitive to specific variability in the gene expression data, both across time points and at time points as reflected in the lack of significant overlap in the genes selected by all of these methods.

The genes observed to be differentially expressed were validated by comparing them to genes that are known to be associated with the diseases. Owing to the fact that different methods apply different algorithms and criteria, the lack of overlap among some methods has been observed and hybrid multiple criteria may help to improve the resulting informative gene sets with both low false-positive rate and good precision, while maintaining sensitivity to relatively subtle changes. This is a subject of our ongoing study as we attempt to resolve the inconsistencies of various filtering methods in order to discover the novel genes that are truly responsible for specific diseases.

The results from our discussed methods not only provide the most probable set of significant genes responsible for the observed profiles and changes during the time courses, but also the ‘stochastic profiles’ of each condition, which we referred to as trajectory analysis. This has been further utilized to discover gene–gene dependence,
for understanding the changes occurring from time to time under drug treatment. We have been able to isolate three clusters of genes whose regulations appear to be inter-dependent. Further investigation of these three clusters is necessary in order to better understand the true nature of the dependence of genes and how this information could be used in treatment of MS. Future work will also include analysis of the genes selected by each method in order to elucidate their biological roles. The three clusters of gene response identified—early, intermediate and late—will be further examined for possible interdependence in biological pathways.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Weinstock-Guttman and Dr Murali Ramanathan for providing the data. This work was supported in part by the National Institute of General Medical Sciences Grant (no. 5P20GM67650-02).

REFERENCES


