Statistics and Probability Letters 79 (2009) 1146–1154

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Gene hunting with forests for multigroup time course data

Ariadni Papana, Hemant Ishwaran [*]

Cleveland State University, United States
Cleveland Clinic, United States

ARTICLE INFO

Article history:
Received 14 July 2008
Received in revised form 7 November 2008
Accepted 1 January 2009
Available online 14 January 2009

ABSTRACT

Gene hunting with forests is a new method for identifying differential gene expression profiles across experimental groups using time course data. Our approach utilizes a multi-dimensional filter that captures the functional nature of the data while adjusting for additional variables that may be part of the experimental design. The filter comprises one component measuring gene profile differences, and another component measuring estimation error. Interesting genes are those having substantial gene profile differences and low estimation error. We refer to this as our Gene Hunting Principle. We illustrate this methodology using a balanced design, involving the effects of muscle group-specific gene expressions on postnatal development. We also consider a more complex experimental design focusing on the effects of aging in the human kidney.

1. Introduction

Gene expression data collected over time, often referred to as time course data, are used to study developmental changes in organisms. As opposed to a static snapshot of the genome obtained when microarray data is collected only at a single time point, time course data provides scientists with richer information and the potential for greater insight into biological processes. By far the most popular approaches to time-course analysis are those based on clustering. Among methods that have been tried are hierarchical clustering (Spellman et al., 1998; Eisen et al., 1998), Bayesian model-based clustering (Ramoni et al., 2002; Zhou et al., 2003), singular value decomposition and principal components (Alter et al., 2000, 2003; Leng and Müller, 2006), hidden Markov models (Schliep et al., 2003), spline-based models (Luan and Li, 2003), and simulated annealing (Lukashin and Fuchs, 2001).

Much of the literature has focused on characterizing gene temporal profiles when data is collected from a single biological group. Much less work, however, has looked at identifying time profiles differences when data is collected from two or more biological groups — what we refer to as multigroup time course data. Among approaches that have been used are hypothesis testing (Bar-Joseph et al., 2003), quadratic regression (Xu et al., 2002), spline-based regression using false discovery rate (FDR) control (Storey et al., 2005), and least-squares parametric variable selection (Conesa et al., 2006). See Bar-Joseph (2004) for a comprehensive review of time-course analysis.

Analyzing time-course microarray data is challenging when there are multiple biological groups involved. In addition to having to deal with the functional nature of the data, one has to contend with the possibility that time profiles may differ across groups, and these profiles may vary by gene as well. Unless there is strong biological theory to guide profile discovery, methodology must be data adaptive and flexible enough to estimate group-temporal patterns without supervision. Relying on pre-specified time profiles (Peddada et al., 2003; Ernst et al., 2005) is unlikely to be successful.

[*] Corresponding author at: Cleveland Clinic, Department of Quantitative Health Sciences, Desk JJN3-01, 9500 Euclid Avenue, Cleveland Ohio, 44195, United States.
E-mail address: hemant.ishwaran@gmail.com (H. Ishwaran).

0167-7152/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2009.01.003

In this paper, we propose a general method for identifying time profile differences from multigroup time-course data. Our approach is based on random forests (Breiman, 2001) and is nonparametric, data-adaptive, and applicable to fairly complex experimental designs. In this approach, the expression data are treated as curves, and a multi-dimensional filter is constructed by combining two measures: a measure of functional information, and a measure of predictiveness for the curves. A Gene Hunting Principle (GHP) that synthesizes this information is then used to identify interesting genes (Section 2). Two different examples, of increasing complexity, are used to illustrate the methodology (Sections 3 and 4).

2. Compressing the functional data

Consider microarray gene expression data collected over time for two or more biological groups. Our goal is to identify those genes with different time profiles. Let

$$\Theta = (\Delta_1, \ldots, \Delta_M, \Psi_1, \ldots, \Psi_G)$$

be a multi-dimensional filter calculated from the data, where Δ_m is a measure of functional information, $m = 1, \ldots, M$, and Ψ_g is a measure of predictiveness of the underlying time-profile curves, $g = 1, \ldots, G$. The measures Δ_m and Ψ_g, and parameters M and G, depend upon the nature of the experimental design and are context specific.

In application, the multi-dimensional filter Θ is computed for each gene, but for notational clarity we suppress this dependence throughout the manuscript. Our GHP is stated as follows.

Gene hunting principle. *The filter Θ is computed for each gene, and a set of significant genes is selected based on:*

(A) *Maximization of the functional measure* $(\Delta_1, \ldots, \Delta_M)$.
(B) *Minimization of the predictiveness measure* $(\Psi_1, \ldots, \Psi_G)$.

To illustrate, consider the simplest setting where we have two biological groups (this is similar to our first example to be discussed shortly). For each gene, and each group, a smoothed time profile curve is calculated. Define Δ_1 to be the distance between the two curves, and set $\Psi_g, g = 1, 2$, to be the model error for each curve.

In this example, $M = 1$ and $G = 2$, and the filter is $\Theta = (\Delta_1, \Psi_1, \Psi_2)$. The GHP searches for those genes with substantially different time profiles (i.e., Δ_1 is large) and with underlying curves having small model error (i.e., Ψ_g is small). The premise being that differing time profiles identify interesting genes, but *only* if the underlying curves are to be trusted.

Obviously, the success of our method hinges strongly on the accuracy of the estimated curves. For this reason, we rely on random forests (Breiman, 2001). As we will show, forests will allow us to accurately estimate time profile curves. Furthermore, we will show how to exploit internal forest error measurements, such as variable importance measures, to extend the GHP to more complex experimental designs.

3. Muscle group-specific gene time-profiling: A balanced experimental design

As our first illustration, we consider the microarray data studied in Cheng et al. (2004). The data comprises expression values from extraocular (EOM) and hindlimb rat muscles harvested at birth (day 0) and during postnatal development (days 7, 14, 21, 28, and 45). By sacrificing animals, three independent replicates were obtained at each time point, for each muscle group. Sampled tissues were queried using the Affymetrix RG-U34A platform (8799 probe sets) and then background corrected, normalized, and summarized with MAS 5.0 software. The data are available at the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) data repository under series record accession number GSE903. See www.ncbi.nlm.nih.gov/geo and Cheng et al. (2004) for more information.

The data was used in Cheng et al. (2004) to compare EOM to hindlimb muscles, to provide insight into certain metabolic and neuromuscular diseases. We can cast this problem within our framework as follows. Assume the expression values for each gene in the EOM group ($g = 1$) and hindlimb muscle group ($g = 2$) are of the form:

$$\mathbf{Y}_g = \mathbf{f}_g + \boldsymbol{\epsilon}_g, \quad g = 1, 2,$$

where $\mathbf{Y}_g$ is the vector comprised of the n_g expression values for the g-th group, $\mathbf{f}_g = (f_{g,1}, \ldots, f_{g,n_g})^{\mathrm{T}}$ are the true time-profiles, and $\boldsymbol{\epsilon}_g$ are independent random errors, such that $\mathbb{E}(\boldsymbol{\epsilon}_g) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}_g) = \sigma_g^2\mathbf{I}$. Note that the assumption of independence is reasonable because of the way the data was harvested.

Let $\hat{\mathbf{f}}_g$ be an estimate for $\mathbf{f}_g$. In this example, gene expressions are collected over the same time points for each group. Thus, we can directly compare $\hat{\mathbf{f}}_1$ to $\hat{\mathbf{f}}_2$ on a coordinate-by-coordinate basis. Define,

$$\Delta_1 = \|\hat{\mathbf{f}}_1 - \hat{\mathbf{f}}_2\|_2^2 \quad \text{and} \quad \Psi_g = \mathbb{E}(\|\mathbf{f}_g - \hat{\mathbf{f}}_g\|_2^2), \quad g = 1, 2,$$

where $\|.\|_2$ is the ℓ_2-norm. Note that Ψ_g is the model error for $\hat{\mathbf{f}}_g$, which we shall denote by MER_g.

As in Section 2, our GHP is based on the filter $\Theta = (\Delta_1, \Psi_1, \Psi_2)$. Now we explain an important point regarding our measure of predictiveness, Ψ_g. Although it is common to measure prediction performance using prediction error, our Θ filter uses model error instead. The reason for doing this is that prediction error is confounded with gene-specific variation, whereas model error measures pure error. To see this, let $\mathbf{Y}_g^{\text{new}}$ be a vector of new independent expression values for each group, such that

$$\mathbf{Y}_g^{\text{new}} = \mathbf{f}_g + \boldsymbol{\epsilon}_g^{\text{new}}, \quad g = 1, 2,$$

where $\mathbb{E}(\boldsymbol{\epsilon}_g^{\text{new}}) = 0$ and $\text{Var}(\boldsymbol{\epsilon}_g^{\text{new}}) = \sigma_g^2\mathbf{I}$. The prediction error for $\hat{\mathbf{f}}_g$ is defined as

$$\text{PE}_g := \mathbb{E}(\|\mathbf{Y}_g^{\text{new}} - \hat{\mathbf{f}}_g\|_2^2) = n_g\sigma_g^2 + \mathbb{E}(\|\mathbf{f}_g - \hat{\mathbf{f}}_g\|_2^2) = n_g\sigma_g^2 + \text{MER}_g.$$

Thus, if we use PE_g in place of MER_g for Ψ_g, then minimizing Ψ_g is confounded with the noise of the gene expression data, $n\sigma_g^2$. Minimizing MER_g, on the other hand, goes after pure error, and we are left with genes with true signal differences. This can be stated formally as the following theorem.

Theorem 1. *If the model errors for $\hat{\mathbf{f}}_1$ and $\hat{\mathbf{f}}_2$ are zero, then $\mathbb{E}(\Delta_1) = \|\mathbf{f}_1 - \mathbf{f}_2\|_2^2$. That is, the mean distance-squared between curves is equal to the distance-squared between their true time-profiles.*

Proof. Expanding Δ_1 gives,

$$\begin{aligned}\Delta_1 &= \|(\hat{\mathbf{f}}_1 - \mathbf{f}_1) - (\hat{\mathbf{f}}_2 - \mathbf{f}_2) + (\mathbf{f}_1 - \mathbf{f}_2)\|_2^2 \\ &= \|\hat{\mathbf{f}}_1 - \mathbf{f}_1\|_2^2 + \|\hat{\mathbf{f}}_2 - \mathbf{f}_2\|_2^2 + \|\mathbf{f}_1 - \mathbf{f}_2\|_2^2 - 2(\hat{\mathbf{f}}_1 - \mathbf{f}_1)^{\text{T}}(\hat{\mathbf{f}}_2 - \mathbf{f}_2) \\ &\quad + 2(\hat{\mathbf{f}}_1 - \mathbf{f}_1)^{\text{T}}(\mathbf{f}_1 - \mathbf{f}_2) - 2(\hat{\mathbf{f}}_2 - \mathbf{f}_2)^{\text{T}}(\mathbf{f}_1 - \mathbf{f}_2).\end{aligned}$$

Taking expectations,

$$\begin{aligned}\mathbb{E}(\Delta_1) &= \text{MER}_1 + \text{MER}_2 + \|\mathbf{f}_1 - \mathbf{f}_2\|_2^2 - 2\,\mathbb{E}\left\{(\hat{\mathbf{f}}_1 - \mathbf{f}_1)^{\text{T}}(\hat{\mathbf{f}}_2 - \mathbf{f}_2)\right\} \\ &\quad + 2\,\mathbb{E}\left\{(\hat{\mathbf{f}}_1 - \mathbf{f}_1)^{\text{T}}(\mathbf{f}_1 - \mathbf{f}_2)\right\} - 2\,\mathbb{E}\left\{(\hat{\mathbf{f}}_2 - \mathbf{f}_2)^{\text{T}}(\mathbf{f}_1 - \mathbf{f}_2)\right\}.\end{aligned}$$

Let $\delta_{g,i} = f_{g,i} - \hat{f}_{g,i}$. If $\text{MER}_g = 0$, then

$$0 = \mathbb{E}\left(\sum_{i=1}^{n_g} \delta_{g,i}^2\right) = \sum_{i=1}^{n_g} \mathbb{E}(\delta_{g,i}^2).$$

Therefore, $\delta_{g,i} = 0$ a.e.[$\mathbb{P}$]. Consequently all terms on the right-hand side of $\mathbb{E}(\Delta_1)$ cancel, excepting $\|\mathbf{f}_1 - \mathbf{f}_2\|_2^2$. □

3.1. Results

Predicted gene expressions $\hat{\mathbf{f}}_g$, and prediction errors $\widehat{\text{PE}}_g$, were estimated independently for each gene and each muscle group using random forest regression (Breiman, 2001). Time was used as the predictor, and expression values as the response. Computations were implemented using the randomForest R-package (Liaw and Wiener, 2002). In each case, 1000 trees were grown with all software parameters set to default settings. To estimate σ_g^2 we used leave-one-out cross-validation. For each gene, an orthogonal polynomial model was fit using least squares to data for group g using time as the predictor. The optimal degree for the polynomial was determined by minimizing cross-validated prediction error. The optimal polynomial model was then used to estimate σ_g^2 (using adjusted mean square error from least squares fitting). Denoting this estimator by $\hat{\sigma}_g^2$, we estimated model error by

$$\Psi_g = \max(\widehat{\text{PE}}_g - n_g\hat{\sigma}_g^2, 0), \tag{1}$$

(We note only 1.21% and 1.69% of the 8799 probe sets were found to have negative values $\widehat{\text{PE}}_g - n_g\hat{\sigma}_g^2$ for the EOM and hindlimb muscle groups, respectively).

Applying the GHP, we found 170 significant genes when selecting those genes with model errors smaller than their 80th percentile, and having Δ_1 values larger than their 90th percentile. Note that an 80th percentile cut-off might seem high for the model error, but because the distribution of Θ was highly skewed, we found that without using a large percentile for model error, gene lists were too sparse. Additional information is given in Table 1. Note that Table 1 also lists results using the filter $\Theta^* = (\Delta_1^*, \Psi_1, \Psi_2)$, where $\Delta_1^* = \|\hat{\mathbf{f}}_1 - \hat{\mathbf{f}}_2\|_1$ and $\|\cdot\|_1$ is the ℓ_1-norm. Using an ℓ_1-based measure has the potential to be more robust.

Table 1 showed very little difference between Θ and Θ^*, but to ensure robustness we focused on Θ^* hereafter. Fig. 1 plots Θ^* for the two muscle groups (model error was transformed by taking logs). Similar patterns are seen for both tissues with slightly larger values seen for EOM data. The top 12 time profiles, corresponding to genes with the highest Δ_1^* values, are plotted in Fig. 2. The profiles are clearly different across muscle groups.

The set of Θ^* and Θ-significant genes (using cut-off$_1$; see Table 1) were studied in terms of their functional ontologies as defined by the Gene Ontology (GO) Consortium. There are 3 ontologies defined by the GO Consortium: Biological Process

Table 1
Number of significant genes for cut-off$_1$ = (90%, 80%, 80%), cut-off$_2$ = (90%, 90%, 90%), and cut-off$_3$ = (90%, 99.7%, 99.7%) percentile values for Θ and Θ^*; that is genes with Δ_1 or Δ_1^* values greater than their 90th percentile and with model errors MER$_1$, MER$_2$ smaller than the 80th, 90th, or 99.7th percentile, respectively.

	cut-off$_1$	cut-off$_2$	cut-off$_3$
Θ	170	340	847
Θ^*	179	363	847

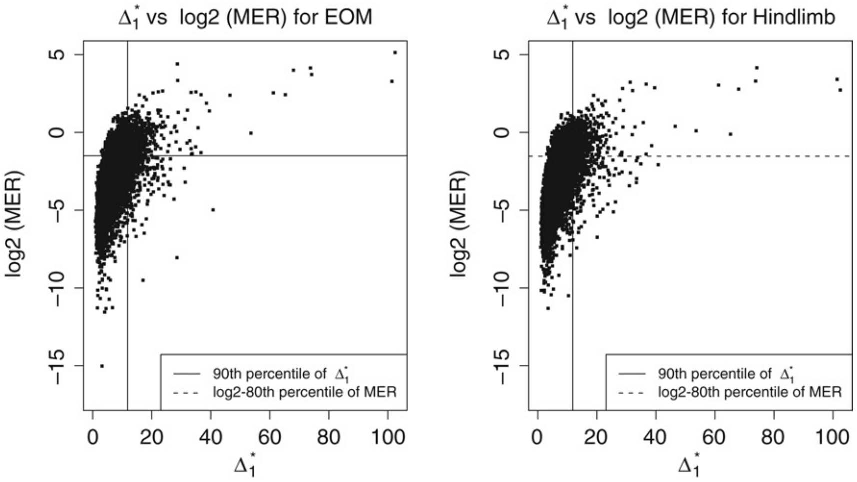

Fig. 1. Plot of Θ^*: Δ_1^* versus log2-transformed model errors for the EOM and hindlimb muscles groups (left and right panels, respectively). The vertical and horizontal lines in the two panels correspond to the 90th percentile of Δ_1^* and the 80th percentile of the log-transformed Ψ_g.

Table 2
Number of identified Biological Processes (BP), Molecular Functions (MF), and Cellular Component (CC) ontologies. Number of genes indicated in parentheses.

	BP	MF	CC
Θ^* (179)	134	132	47
Θ (170)	126	119	48
All genes (8799)	1004	1294	278

(BP); Molecular Function (MF); and Cellular Component (CC). The BP of a gene product is a biological objective to which the gene product contributes and involves the transformation of a physical thing; the MF is what the gene product does at a biochemical level; and the CC is a component of a cell that is part of a larger object or structure (see www.geneontology.org for more details). A break-down of ontologies for our significant genes, as well as the entire set of genes on the array, is given in Table 2. In total there were 134 biological process, 132 molecular functions, and 47 cellular components associated with our 179 Θ^*-significant genes.

The percentage of genes that fall in the most frequented BP, MF, and CC ontologies (with respect to the set of Θ^*-significant genes) are graphed in Fig. 3. Frequencies for Θ and Θ^* are in close agreement. What is most interesting is the difference in frequencies for significant genes to all genes on the array. This shows that selected genes have a different ontology distribution and provides clear evidence that time profiles for significant genes are different over muscle type.

4. Gene hunting for general experimental designs

Now we show how the GHP can be applied to more complex experimental designs. We look at the data in Rodwell et al. (2004) who studied the effects of aging on the human kidney. This data comprises gene expressions from 72 patients, with ages ranging from 27 to 92 years. Kidney samples were obtained from all patients, and these were dissected into 72 cortex samples and 62 medulla samples. Total RNA was isolated from each of the $n = 134$ samples and queried using Affymetrix HG-U133A and HG-U133B high density oligonucleotide arrays (44,928 probesets). MAS 5.0 software was used to normalize the data. The data are available at the Stanford Microarray Database (http://genome-www5.stanford.edu).

Although the data is cross-sectional we can treat it as if it were time-course data by using age as the time variable. This creates an unbalanced design. On top of this, an added wrinkle was that there were additional patient variables that could be included in the analysis. This included gender, race, age, blood pressure, pathology, medications, serum creatinine, and urinary protein concentration data.

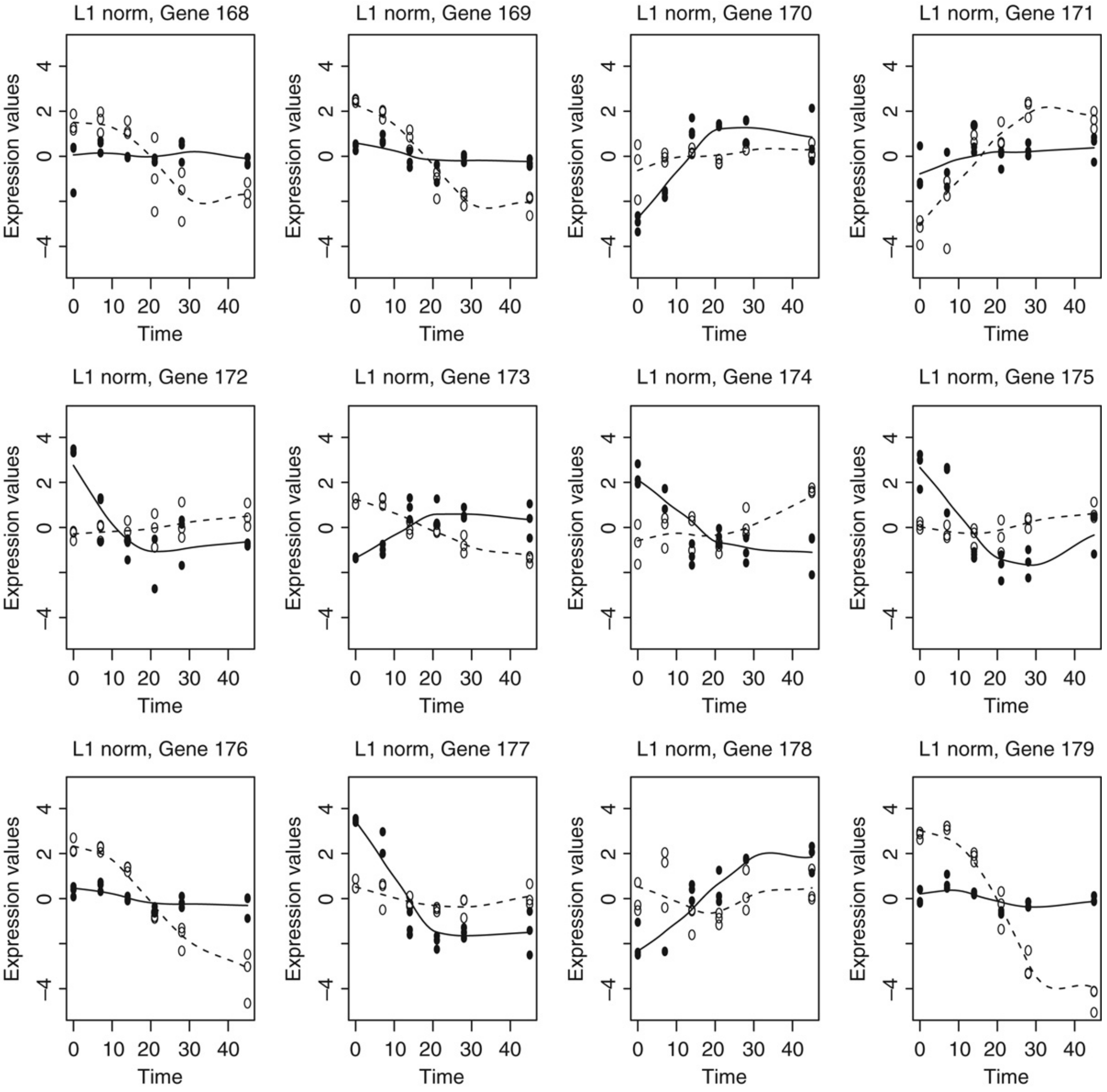

Fig. 2. Top 12 time profiles with the largest Δ_1^* values (Δ_1^* decreases from top to bottom). Expression values and their smoothed loess curves (Cleveland, 1979) are plotted from day 0 through 45. Solid points and lines represent EOM data, whereas empty-circle points and dashed lines are hindlimb data.

To cast this within our framework, we shall assume that the expression value for the i-th tissue sample, for a given gene, is of the form

$$Y_i = f(x_{i,1}, \ldots, x_{i,p}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

where $x_{i,j}$ is sample i's value for the j-th variable, $j = 1, \ldots, p$, and ε_i are independent random errors such that $\mathbb{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. The assumption of independence is again warranted here. It is quite reasonable to assume that the $n = 134$ cortex and medulla tissues are independent, even though in some cases these samples were obtained from the same patient.

The expression values are a function of p variables and thus the dimension of the functional measure is $M = p$. It is unlikely that age dependent gene expressions are affected by many of the factors that were available (Storey et al., 2005; Rodwell et al., 2004; Higgins et al., 2004). Thus, as in previous analyses, we considered only gender and tissue type information in addition to age; so that $p = 3$. The predictiveness component for Θ has dimension $G = 1$, because there is only one model fit for each gene. We fit (2) using random forest regression and define Θ by:

$$\Delta_m = I_m, \quad \text{for } m = 1, \ldots, p, \quad \text{and} \quad \Psi_1 = \mathrm{MER}.$$

Here, I_m are random forest variable importance (VIMP) values. VIMP measures the change in prediction error on a new test case if a variable were removed from the analysis (Breiman, 2001; Ishwaran, 2007). A large positive VIMP indicates predictiveness of a variable, adjusting for remaining variables.

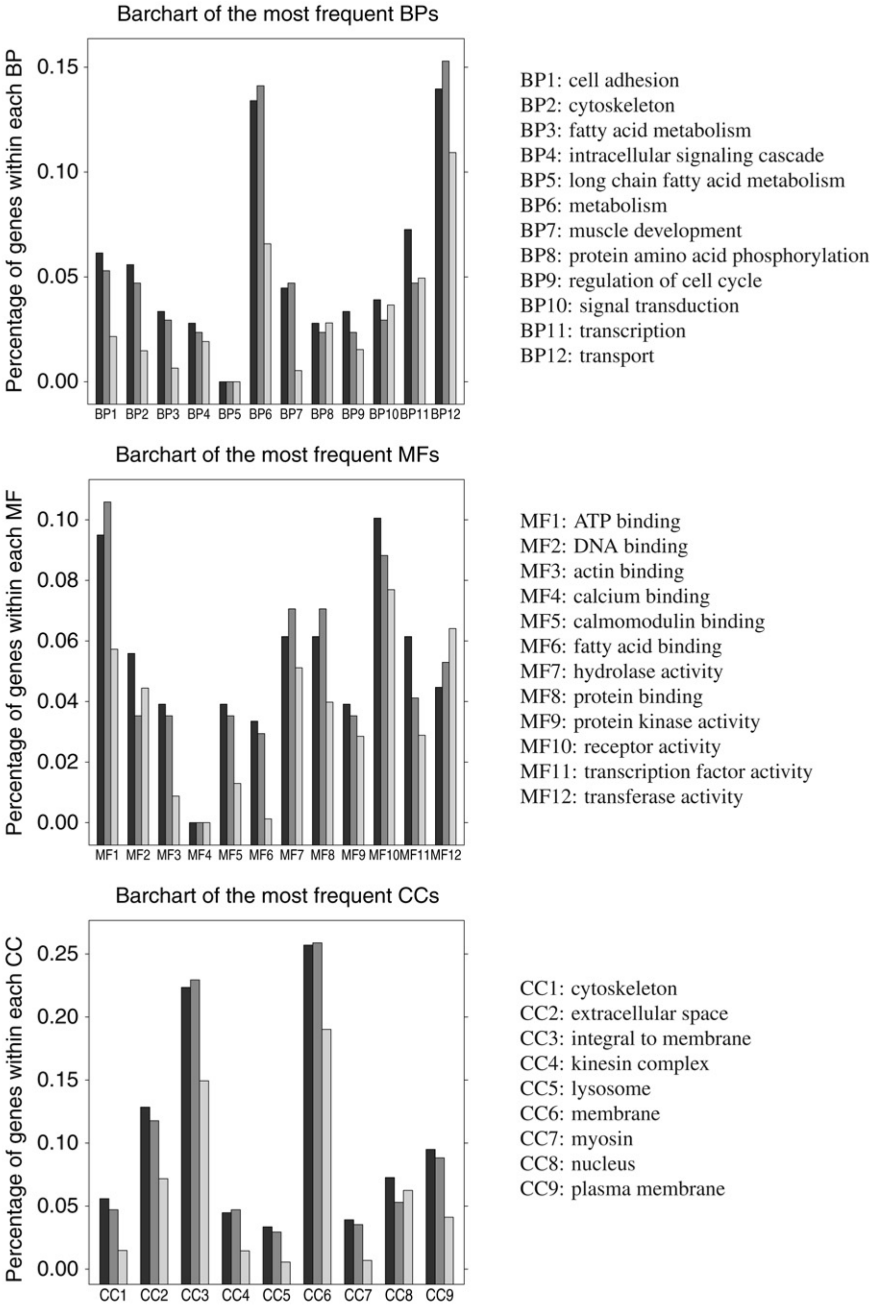

Fig. 3. Percentage of genes with most frequent Biological Processes (BP), Molecular Functions (MF), and Cellular Components (CC). Dark gray and medium gray bars are Θ^* and Θ significant genes. Light gray bars indicate all genes on the array.

The GHP here is quite interesting: find genes with large positive importance values and small model errors.

Remark 1. Note that to ensure proper comparison of VIMP across genes, we scaled I_m by dividing it by the variance of Y.

4.1. Results

Random forest regression was applied to each gene using age, gender, and tissue type as predictors. Each forest comprised 1000 trees, with computations implemented using the randomForest R-package (Liaw and Wiener, 2002) under default settings. Model error was calculated as in (1). For the estimate of σ^2 we used adjusted mean square error from a least squares fit using a linear model with B-splines used for age, and main effects, and interactions included for tissue type and gender.

Table 3
Number of significant genes using threshold values of cut-off$_1$ = (0.1, 0.001, 0.001, 0.01), cut-off$_2$ = (0.01, 0.01, 0.01, 0.01), cut-off$_3$ = (0.005, 0.005, 0.005, 0.01), cut-off$_4$ = (0.002, 0.002, 0.002, 0.01), and cut-off$_5$ = (0.001, 0.001, 0.001, 0.01) for Θ.

Θ cut-off	Number of significant genes
cut-off$_1$	36
cut-off$_2$	96
cut-off$_3$	209
cut-off$_4$	326
cut-off$_5$	378

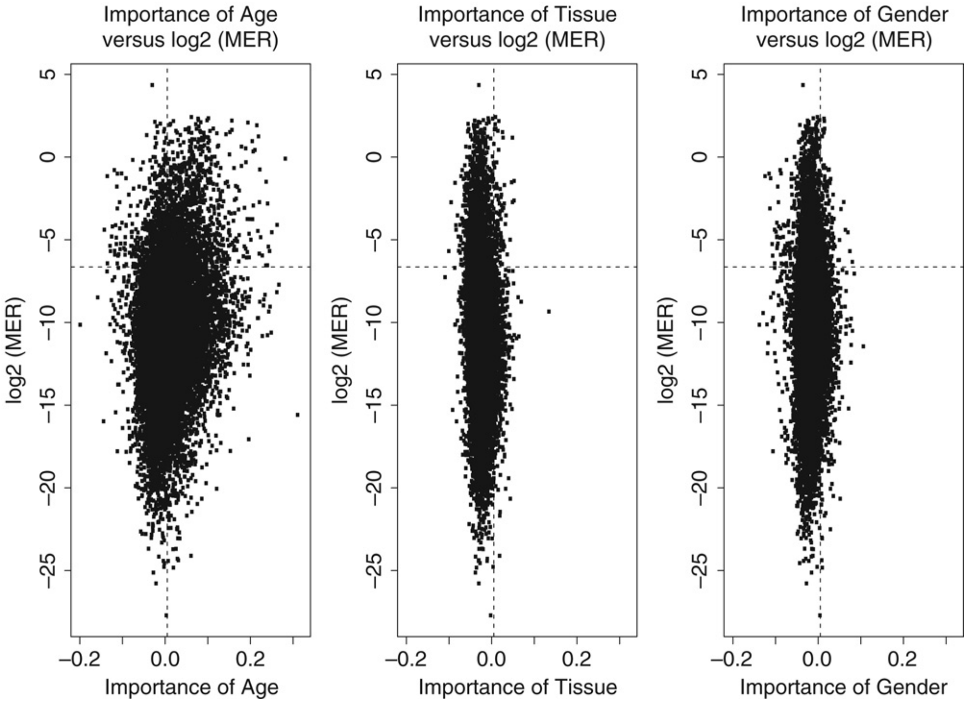

Fig. 4. Θ values: VIMP for age, tissue, and gender versus log2-transformed model error (left, middle, and right panels). The vertical and horizontal lines in each plot correspond to the Θcut-off$_3$ = (0.005, 0.005, 0.005, 0.01).

Table 3 shows how the number of significant genes vary as a function of thresholding values applied to $\Theta = (I_{\text{age}}, I_{\text{tissue}}, I_{\text{gender}}, \text{MER})$. Significant genes were annotated using GO methodology, but this data is suppressed for brevity.

Fig. 4 plots Θ values from the analysis. VIMP for I_{age}, I_{tissue}, and I_{gender} are plotted versus log-transformed model error. Similar patterns are seen in the three plots, however VIMP for age is systematically larger—thus revealing a strong age effect. Fig. 5 plots VIMP for age versus tissue and age versus gender (left and right panels, respectively). The time profiles corresponding to the gene with the highest I_{tissue} value out of the significant set of genes, are plotted in Fig. 6. The profile shows a very interesting age-gender-tissue interaction.

5. Discussion

Gene hunting with random forests is a new method that can be used to find gene expression time-profile differences across biological groups. It can be used for fairly general experimental designs, even those that are unbalanced and that may involve additional experimental variables. Because the approach is based on random forest methodology, it is nonparametric and data adaptive. Furthermore, it can be computed efficiently, even for large microarray studies, and is relatively easy to use.

At the same time, as this is a new methodology, some interesting unanswered questions have emerged. One key issue is the assumption of independence across time used in both our examples. Although this is an assumption that can often hold in animal based experiments (as in Section 3), it brings up the question of whether the methodology can be applied to general time course settings where dependence across time may be at play. Looking carefully back at our examples we see that independence was needed primarily to get estimates for the model error for our predictors (via subtracting estimates of the internal noise σ^2 from the estimated prediction error). This naturally suggests dependent settings could be handled if model error can be estimated without requiring independence. This is an interesting area for future research.

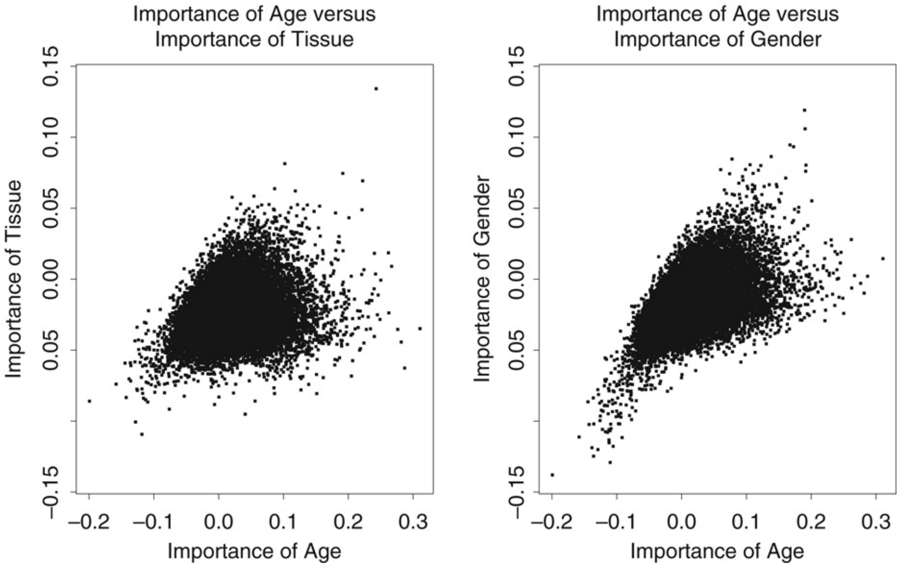

Fig. 5. VIMP for age versus tissue, and age versus gender (left and right panels, respectively).

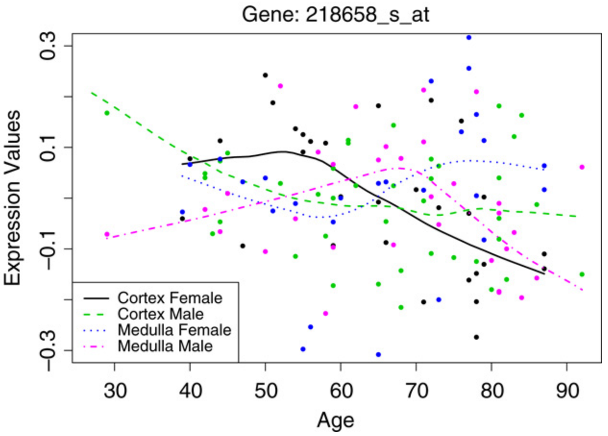

Fig. 6. Time profiles for gene with highest I_{tissue} value from set of Θ-significant genes defined by cut-off$_3$ (Table 3).

Another issue has to do with thresholding. There is now a vast literature in the field of bioinformatics that deals with thresholding p-values for gene expression data (for example, see Datta and Datta (2005)). However, we have taken a prediction approach where there is much less known. The issue of how to appropriately threshold VIMP from forests is a new paradigm where almost nothing seems to be known. Future research in this area promises to be exciting and rewarding.

References

Alter, O., Brown, P.O., Botstein, D., 2000. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. USA 97 (18), 10101–10106.

Alter, O., Brown, P.O., Botstein, D., 2003. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. Proc. Natl. Acad. Sci. USA 100 (6), 3351–3356.

Bar-Joseph, Z., 2004. Analyzing time series gene expression data. Bioinformatics 20 (16), 2493–2503.

Bar-Joseph, Z., Gerber, G., Simon, I., Gifford, D.K., Jaakkola, T.S., 2003. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. Proc. Natl. Acad. Sci. USA 100 (18), 10146–10151.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Cheng, G., Merriam, A.P., Gong, B., Leahy, P., Khanna, S., Porter, J.D., 2004. Conserved and muscle-group-specific gene expression patterns shape postnatal development of the novel extraocular muscle phenotype. Physiol. Genom. 18, 184–195.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. J. Amer. Stat. Assoc. 74 (368), 829–836.

Conesa, A., Nueda, M.J., Ferrer, A., Talon, M., 2006. maSigPro: A method to identify significantly differential expression profiles in time-course microarray experiments.. Bioinformatics 22 (9), 1096–1102.

Datta, S., Datta, S., 2005. Empirical Bayes screening (EBS) of many p-values with applications to microarray studies. Bioinformatics 21, 1987–1994.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95, 14863–14868.

Ernst, J., Nau, G.J., Bar-Joseph, Z., 2005. Clustering short time series gene expression data. Bioinformatics 21 (Suppl.1), i159–i168.

Higgins, J.P., Wang, L., Kambham, N., Montgomery, K., Mason, V., Vogelmann, S.U., Lemley, K.V., Brown, P.O., Brooks, J.D., van de Rijn, M., 2004. Gene expression in the normal adult human kidney assessed by complementary DNA microarray. Mol. Biol. Cell 15 (2), 649–656.

Ishwaran, H., 2007. Variable importance in binary regression trees and forests. Elec. J. Stat. 1, 519–537.

Leng, X., Müller, H.G., 2006. Classification using functional data analysis for temporal gene expression data. Bioinformatics 22 (1), 68–76.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22.

Luan, Y., Li, H., 2003. Clustering of time-course gene expression data using a mixed-effects model with B-splines. Bioinformatics 19 (4), 474–482.

Lukashin, A.V., Fuchs, R., 2001. Analysis of temporal gene expression profiles: Clustering by simulated annealing and determining the optimal number of clusters. Bioinformatics 17 (5), 405–414.

Peddada, S.D., Lobenhofer, E.K., Li, L., Afshari, C.A., Weinberg, C.R., Umbach, D.M., 2003. Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. Bioinformatics 19 (7), 834–841.

Ramoni, M.F., Sebastiani, P., Kohane, I.S., 2002. Cluster analysis of gene expression dynamics. Proc. Natl. Acad. Sci. USA 99 (14), 9121–9126.

Rodwell, G.E.J., Sonu, R., Zahn, J.M., Lund, J., Wilhelmy, J., Wang, L., Xiao, W., Mindrinos, M., Crane, E., Segal, E., Myers, B.D., Brooks, J.D., Davis, R.W., Higgins, J., Owen, A.B., Kim, S.K., 2004. A transcriptional profile of aging in the human kidney. PLOS Bio. 2 (12), 2191–2201.

Schliep, A., Schönhuth, A., Steinhoff, C., 2003. Using hidden Markov models to analyze gene expression time course data. Bioinformatics 19 (Suppl.1), i255–263.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Mol. Bio. Cell 9, 3273–3297.

Storey, J.D., Xiao, W., Leek, J.T., Tompkins, R.G., Davis, R.W., 2005. Significance analysis of time course microarray experiments. Proc. Natl. Acad. Sci. USA 102 (36), 12837–12842.

Xu, X.L., Olson, J.M., Zhao, L.P., 2002. A regression-based method to identify differentially expressed genes in microarray time course studies and its applications in an inducible hunington's disease transgenic model. Human Mol. Gen. 11 (17), 1977–1985.

Zhou, C., Wakefield, J.C., Self, S.G., 2003. Modelling gene expression data over time: Curve clustering with informative prior distributions. Bayesian Stat. 7, 721–732.