Identifying Differentially Expressed Genes in Unreplicated Multiple-Treatment Microarray Timecourse Experiments

Rhonda DeCook  
*Iowa State University*

Dan Nettleton  
*Iowa State University*, dnett@iastate.edu

Carol Foster  
*Iowa State University*

Eve S. Wurtele  
*Iowa State University*, mash@iastate.edu

Follow this and additional works at: [http://lib.dr.iastate.edu/stat_las_preprints](http://lib.dr.iastate.edu/stat_las_preprints)

Part of the *Statistics and Probability Commons*

Recommended Citation
DeCook, Rhonda; Nettleton, Dan; Foster, Carol; and Wurtele, Eve S., "Identifying Differentially Expressed Genes in Unreplicated Multiple-Treatment Microarray Timecourse Experiments" (2004). *Statistics Preprints*. 94.  
[http://lib.dr.iastate.edu/stat_las_preprints/94](http://lib.dr.iastate.edu/stat_las_preprints/94)

This Article is brought to you for free and open access by the Statistics at Iowa State University Digital Repository. It has been accepted for inclusion in Statistics Preprints by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Identifying Differentially Expressed Genes in Unreplicated Multiple-Treatment Microarray Timecourse Experiments

Abstract
Microarray technology has become widespread as a means to investigate gene function and metabolic pathways in an organism. A common experiment involves probing, at each of several time points, the gene expression of experimental units subjected to different treatments. Due to the high cost of microarrays, such experiments may be performed without replication and therefore provide a gene expression measurement of only one experimental unit for each combination of treatment and time point. Though an experiment with replication would provide more powerful conclusions, it is still possible to identify differentially expressed genes and to estimate the number of false positives for specified rejection region when the data is unreplicated. We present a method for identifying differentially expressed genes in this situation that utilizes polynomial regression models to approximate underlying expression patterns. In the first stage of a two-stage permutation approach, we choose a 'best' model at each gene after considering all possible regression models involving treatment effects, terms polynomial in time, and interactions between treatments and polynomial terms. In the second stage, we identify genes whose 'best' model differs significantly from the overall mean model as differentially expressed. The number of expected false positives in the chosen rejection region and the overall proportion of differentially expressed genes are both estimated using a method presented by Storey and Tibshirani (2003, Proceedings of the National Academy of Sciences 100, 9440-9445). For illustration, the proposed method is applied to an Arabidopsis thaliana microarray data set.

Keywords
differentially expressed genes, microarrays, unreplicated data

Disciplines
Statistics and Probability

Comments
Identifying Differentially Expressed Genes in Unreplicated Multiple-Treatment Microarray Timecourse Experiments

Rhonda DeCook,1 Dan Nettleton,1 Carol Foster,2 and Eve S. Wurtele2

Departments of 1Statistics and 2Botany, Iowa State University, Ames, Iowa 50011-1210, U.S.A.

*email: rdecook@iastate.edu

Summary. Microarray technology has become widespread as a means to investigate gene function and metabolic pathways in an organism. A common experiment involves probing, at each of several time points, the gene expression of experimental units subjected to different treatments. Due to the high cost of microarrays, such experiments may be performed without replication and therefore provide a gene expression measurement of only one experimental unit for each combination of treatment and time point. Though an experiment with replication would provide more powerful conclusions, it is still possible to identify differentially expressed genes and to estimate the number of false positives for a specified rejection region when the data is unreplicated. We present a method for identifying differentially expressed genes in this situation that utilizes polynomial regression models to approximate underlying expression patterns. In the first stage of a two-stage permutation approach, we choose a ‘best’ model at each gene after considering all possible regression models involving treatment effects, terms polynomial in time, and interactions between treatments and polynomial
terms. In the second stage, we identify genes whose ‘best’ model differs significantly from
the overall mean model as differentially expressed. The number of expected false positives
in the chosen rejection region and the overall proportion of differentially expressed genes are
both estimated using a method presented by Storey and Tibshirani (2003, Proceedings of
the National Academy of Sciences 100, 9440-9445). For illustration, the proposed method
is applied to an Arabidopsis thaliana microarray data set.

Key words: Differentially expressed genes; Microarrays; Unreplicated data

1 Introduction

Present microarray technology allows a researcher to simultaneously measure the mRNA
expression level of thousands of genes from a single experimental unit. This ability serves
as a powerful genomics tool for uncovering gene function. By applying this technology to a
multiple-treatment timecourse experiment, a researcher can locate genes whose expression
patterns over time are biologically interesting as defined in a number of ways. For example,
genesis of biological interest may include genes whose expression changes significantly over
time in an identical fashion for all treatments (time main effects), genes whose expression
differs significantly among treatments but remains constant over time (treatment main ef-
tects), or genes whose pattern of expression over time differs significantly among treatments
(time-by treatment interaction).

Though the cost of microarrays has greatly decreased since their emergence, a multiple-
treatment microarray timecourse experiment can still be quite costly. Therefore, a researcher interested in such an experiment may choose to perform it with no replications in hopes of finding interesting expression patterns that will be studied more intensively in follow-up experiments. In this situation, a statistical analysis must rely on only one expression measurement per treatment-time combination for each gene, making separation of signal and noise a difficult task.

Even when replication is present, a microarray data set that holds information on both treatment and time effects poses challenges to the statistical researcher that go beyond the one-way ANOVA methods or the cluster analysis methods that are often applied to microarray data. Issues such as normalization, signal-to-noise ratios, and multiple testing are still present, but the fact that numerous alternative hypotheses may be of interest at each of thousands of genes adds a layer of complexity to the analysis.

In a microarray experiment that compares only two groups, the same test statistic, namely a standard \( t \)-test or one of its modifications, can be used to test for differential expression at every gene. Storey and Tibshirani (2001) provide a method in this scenario for identifying differentially expressed genes that estimates the positive False Discovery Rate (pFDR) for the chosen rejection region. To apply this method in a multiple-treatment timecourse experiment, a researcher could use the \( F \)-statistic comparing the most complex polynomial regression model with positive degrees of freedom for error to the model that assumes a common mean for all conditions to test for the existence of any effects at each gene. Employing this overall \( F \)-test for all genes, however, would waste degrees of freedom for genes whose true expression patterns can be adequately described with a relatively simple alternative polynomial regression model. An \( F \)-statistic comparing a simpler polynomial regression
model to the overall mean model would have more power for detecting genes with simpler
expression patterns.

We propose a method for locating genes whose expression patterns over time differ in
any way from the overall mean model that provides more power for simpler alternatives than
the overall $F$-test described above. A concurrent flat line profile for all treatment groups
represents the expected expression pattern at any gene with a true null hypothesis. This null
reflects the reasoning that any gene represented by a pattern other than the overall mean
model is potentially biologically interesting. To test this null hypothesis, we first choose a
‘best’ model at each gene from the pool of candidate models including all possible regres-
sion models involving treatment effects, terms polynomial in time, and interactions between
treatments and polynomial terms. The model choice criterion is based on the vector of $F$-
statistics comparing each candidate model to the overall mean model. The model with the
most extreme $F$-statistic when compared to the distribution of similar $F$-statistics gener-
ated from permuted data sets is considered the ‘best’ model. In the second step, we identify
genes whose ‘best’ model differs significantly from the overall mean model by using the same
permuted data sets as those used in the model-selection step to generate a relevant $p$-value
for each gene. A $p$-value threshold is chosen that coincides with both a list of significant
genes and an estimated number of expected false positives among the genes declared to be
significant.

There are several benefits of our proposed procedure. First, it allows us to associate a
gene with a ‘best’ model requiring fewer degrees of freedom than the most complex model
possible. The $F$-statistic comparing this ‘best’ model with the overall mean model will
potentially have more degrees of freedom for error and more power for detecting differen-
tial expression than the overall $F$-test. Furthermore, differentially expressed genes will be automatically sorted into groups of genes exhibiting similar expression patterns. Lastly, considering time as a quantitative factor allows us to detect genes exhibiting treatment-by-time interaction in unreplicated data because the pool of candidate models includes those that fit a separate polynomial to each treatment group while still allowing for degrees of freedom for error.

In the next section, we discuss the form of the data, the model selection process, the test statistic, and the multiple comparison adjustment. Section 3 provides a simulation study comparing sensitivity and specificity for our method and the overall $F$-test method when six particular alternative expression patterns are present in the data. In Section 4, we apply our method to data generated from a multiple-treatment timecourse experiment that exposed three genetic lines of the *Arabidopsis thaliana* plant to five different durations of ultra-violet light. In Section 5 we discuss recent related work in this area and in Section 6 we provide some final discussion.

2 Methods

2.1 Notation and Hypotheses

Suppose an unreplicated multiple-treatment timecourse experiment includes $J$ treatments denoted by $1, 2, \ldots, J$ and $T$ time points denoted by $1, 2, \ldots, T$. This experiment would require $M = J \cdot T$ microarrays each providing an expression level measure on $G$ distinct genes. Though this experiment provides only one expression measurement for a particular
gene-treatment-time combination, \( Y_{gjt} \), it does provide \( M \) expression measurements per gene.

The expression at each gene can be described by a model allowing for a unique expected value at each treatment-time combination and a random error term:

\[
Y_{gjt} = \mu_{gjt} + \epsilon_{gjt} \quad \text{with} \quad \epsilon_{gjt} \sim iid(0, \sigma^2_g).
\]  

As stated in Section 1, we consider any gene with an expression pattern different than the overall mean pattern to be potentially biologically interesting. Using this reasoning, we propose a set of hypotheses to be tested at every gene that has the ability to detect a variety of interesting expression patterns. The mean structure at a gene with a true null hypothesis depends neither on treatment group nor time and can therefore be fully described by the single parameter \( \mu_g \). This leads to the following set of hypotheses:

\[
H_{0g} : \mu_{gjt} = \mu_g \quad \text{versus} \quad H_{1g} : \text{not}H_{0g}
\]  

Hereafter, we use the terms overall mean model and null model interchangeably and use both as a reference to model (1) under \( H_{0g} \).

2.2 Model Selection

In a multiple-treatment microarray timecourse experiment, there are many expression patterns that may be of interest to a researcher. Exploring the data with traditional model selection methods such as BIC or AIC could provide useful information about the existence of particular patterns, but they neither incorporate a formal hypothesis test nor provide a relevant \( p \)-value for differential expression (as defined by \( H_{0g} \) being false).

To provide such a \( p \)-value, we first select a ‘best’ model at each gene from a given candidate pool using a permutation method described below. This selection process provides
us with a test statistic for testing the hypotheses in (2). We compare this test statistic to a permutation distribution of similar test statistics generated under $H_{0g}$ to compute a relevant permutation $p$-value to be used for determining significance of differential expression.

The pool of candidate models consists of all possible regression models involving treatment effects, terms polynomial in time, and interactions between treatments and polynomial terms that still allow for degrees of freedom for error. We exclude the overall mean model from the candidate pool and when a particular term is included in a candidate model, we assume that all lower order terms are also included to coincide with the hierarchical order principle described in Wu and Hamada (2000). Thus, a given experiment with $M$ observations per gene will have $S$ models in the candidate pool.

In the selection process, we first compute the vector of observed $F$-statistics $F_o = (F_{1o}, F_{2o}, \cdots, F_{So})$ comparing each possible model $s$ to the null model at each gene in the original data set. (For simplicity, we will suppress the subscript $g$ in our notation because the same process applies to all genes.) To generate a possible realized vector of $F$-statistics under $H_0$, we permute the labels on the $M$ microarrays and then re-compute the previously described vector of $F$-statistics $F^*_1 = (F^*_{11}, F^*_{21}, \cdots, F^*_{S1})$ at each gene. Creating a permuted data set in this manner preserves the unknown covariance structure between genes in a single experimental unit, while providing a possible realized data set generated under $H_0$. Repeating this process for a total of $B$ permuted data sets randomly chosen from the $M!$ possible sets provides us with a permutation distribution of $F$-statistic vectors $F^*_1, F^*_2, \cdots, F^*_B$ at every gene.

Our model selection method chooses the model whose $F$-statistic is most extreme when compared to the permutation distribution of $F$-statistics computed for the same model. To
make this decision, we compute an approximate permutation \( p_{s_0}^* \) for each model \( s \),

\[
p_{s_0}^* = \frac{\sum_{b=0}^{B} I(F_{s_b}^* \geq F_{s_0})}{B + 1}
\]

where \( F_{s_0}^* \equiv F_{s_0} \), and choose the ‘best’ model \( m_o \in \{1, \ldots, S\} \) as the model with the smallest permutation \( p \)-value, i.e. choose \( m_o \) such that \( p_{m_o}^* = \min_s p_{s_0}^* \). Under the null-hypothesis, each \( p_{s_0}^* \) behaves as a conservative \( p \)-value in that \( P_{H_0}(p_{s_0}^* \leq \alpha) \leq \alpha \) for any \( \alpha \in [0,1] \).

### 2.3 Test Statistic and \( P \)-values

To test the set of hypotheses in (2), we need (i) a relevant test statistic, (ii) knowledge of the test statistic’s distribution under \( H_0 \), and (iii) some general knowledge of the test statistic behavior when \( H_0 \) is false. For (i), we propose using the minimum permutation \( p \)-value \( p_{m_o}^* \in (0,1] \). For (ii), we generate an empirical distribution of test statistics under \( H_0 \) denoted as \( p_{m_1}^*, p_{m_2}^*, \ldots, p_{m_B}^* \) by sending each permuted data set through the same model selection process as the original data. In essence, when the null hypothesis is true, \( p_{m_o}^* \) acts as the minimum order statistic computed from \( S \) dependent Uniform(0,1) random variables. Therefore, we expect the distribution of the test statistic under \( H_0 \) to be stochastically smaller than a Uniform(0,1) distribution, but because we do not know the dependence structure in the \( S \) random variables we use a permutation distribution of the test statistic rather than a theoretical distribution to test the hypotheses of interest. With respect to (iii), when the null hypothesis is false we expect at least one of the values in \( F_o \) to be more extreme than most of its \( F \)-statistic counterparts generated from the permuted data. This would equate to a relatively small test statistic \( p_{m_o}^* \) for the original data when compared to the permutation distribution of test statistics \( p_{m_1}^*, p_{m_2}^*, \ldots, p_{m_B}^* \).
By comparing the observed test statistic to the permutation distribution of test statistics we compute an approximate permutation $p$-value $p^{**}$ appropriate for testing the set of hypotheses in (2) at each gene using the following expression:

$$
p^{**} = \frac{\sum_{b=0}^{B} I(p^*_{m,0} \leq p^*_{m,b})}{B + 1}
$$

where $p^*_{m,0} \equiv p^*_{m,o}$.  

(3)

Under the null-hypothesis, each $p^{**}$ behaves as a conservative $p$-value in that $P_{H_0}(p^{**} \leq \alpha) \leq \alpha$ for any $\alpha \in [0, 1]$. This behavior is demonstrated empirically by the distribution of $p^{**}$ values taken from 114,000 null genes in our simulated data study (Section 3). The proportion of $p^{**}$ values that fell below 0.001, 0.01, 0.05, 0.10 were 0.0010, 0.0099, 0.0510, and 0.1010, respectively.

The set of empirical $p$-values can be adjusted for multiple comparisons and used to determine significance by using one of numerous proposed methods. We have opted to use the method proposed by Storey and Tibshirani (2003). This method estimates the proportion of genes with a true null hypothesis using the fact that the $p$-values generated from null genes are uniformly distributed. Using this estimate, the expected proportion of false positives among all significant genes, also known as the $q$-value, is estimated for each gene (i.e. for each potential significance threshold). A threshold is chosen that coincides with an acceptable number of genes declared to be differentially expressed and an acceptable $q$-value.

Up to this point, we have described two possible methods which could be used for identifying differentially expressed genes in a multiple-treatment timecourse experiment. For ease of comparison in the simulation study, we will refer to our proposed method as the CALM method to emphasize that its candidate pool considers all models. The second method, related to the overall $F$-test, computes only one $F$-statistic at each gene comparing the most
complex model possible to the overall mean model and we will refer to this method as the Overall-$F$ method. We now introduce a third method that follows the same procedure as the CALM method, but considers only those models that include an interaction term in its candidate pool. This method would be most useful for the researcher who is only interested in locating genes whose expression patterns over time differ across treatment groups. We refer to this method as the COIM method to emphasize that its candidate pool considers only interaction models.

3 Simulation Study

We conducted a Monte Carlo simulation study to compare the sensitivity and specificity of the CALM method, the COIM method, and the Overall-$F$ method when six particular alternative gene expression patterns are present in the data. We used 2500 permutations for both the CALM and COIM methods.

3.1 Design

The design of the simulated data is based on the motivating experiment described in Section 4 that explored the gene expression of three treatment groups over five time points with one microarray assigned to each treatment-time combination. For this simulation study, each of the hypothetical microarrays contains 5000 probe sets (or genes), and the five time points are assigned at uniform intervals of 0, 6, 12, 18, and 24 hours.

We chose to include the given six alternative patterns (Table 1) primarily because their mean structures represent potentially biologically interesting phenomena. Beyond this initial
criterion for inclusion, we were also interested in including both simple and complex as well as polynomial and non-polynomial alternatives to permit a comparison of methods in a variety of situations. The mean structures of the first two alternatives (\(i\) & \(ii\)) are considered simple and can be fully described by a polynomial regression model in the CALM candidate pool that does not include a treatment-by-time interaction. The mean structures of the next two alternatives (\(iii\) & \(iv\)) can be described by a candidate model considered by both the CALM and COIM methods. However, these mean structures are more complex and require an interaction term to fully describe their pattern. The last two alternatives (\(v\) & \(vi\)) have mean structures that can not be described by a model in any candidate pool, but may still be of interest to the researcher. Table 1 describes the specific expression pattern used for each alternative and lists the number of parameters required to fully describe the mean structure when a candidate polynomial regression model is applicable.

[Insert Table 1]

To generate each data set, we simulated each alternative at 200 genes by adding a random error term \(\epsilon \sim N(0, 1)\) to each of the 15 expected values dictated by the given alternative. For the first four alternatives, we chose mean structures that would provide a common Type I and Type II error rate when using the relevant \(F\)-statistic to test for differential expression. This relevant \(F\)-statistic uses the simplest model describing the alternative as the full model and the the overall mean model as the reduced model. The chosen mean structures coincide with a comparison-wise Type I error rate of 0.005 and Type II error rate of 0.20 for all four alternatives. With respect to the non-polynomials, we simulated alternative \(v\) by including a jump of 2.5 standard deviations at the time midpoint. The
horizontal asymptotes of the logistic growth curves in alternative $vi$ have heights of 3, 4, and 5 standard deviations, respectively, relative to time point 0, and each logistic curve is within 0.01 standard deviations of its asymptote by the 18 hour time point. We simulated the remaining 3800 genes in each data set as a null model genes by randomly generating 15 values from the $N(0,1)$ distribution. Thus, the null hypothesis is false at 24% of the genes in each simulated data set.

3.2 Results

3.2.1 Receiver Operating Characteristic (ROC) curves

Figure 1 shows the ROC curves for each method and each alternative computed as an average of 30 simulated data sets. We generated each ROC curve using test statistics from the genes simulated as the given alternative and the full set of null model genes. The vertical axis for each plot represents sensitivity, which is the proportion of truly differentially expressed genes that were declared to be significant. The horizontal axis is plotted as 1-specificity, which is the proportion of null model genes declared as differentially expressed. Each point on the curve represents a possible rejection region that coincides with a specific number of genes declared to be significant. ROC curves are commonly plotted from 0 to 1.0 on both axes, but we have graphed the plots with a horizontal axis from 0 to 0.10 for a more discerning view in the rejection regions likely to be of interest in practice.

[Insert Figure 1]
Figure 1. ROC curves for each method at each of the six alternatives $i$ through $vi$ included in the simulated data based on the average of 30 simulated data sets. The solid line, dotted line, and dashed line represent the CALM method, COIM method, and Overall-$F$ method, respectively.

By comparing the ROC curves in Figure 1, we see that the CALM method performs best for the simpler alternatives $i$ & $ii$, and it is followed by the COIM method and Overall-$F$ method, respectively. The curves for these two alternatives also illustrate a general trend that as the alternative polynomial patterns move from simple to complex (i.e. more parameters are required to describe the alternative’s mean structure), the distance between the CALM and COIM method ROC curves decreases until the alternative complexity becomes large enough to require an interaction term for full description, as in alternatives $iii$ & $iv$. Once this level of alternative complexity is reached, the COIM method overtakes the CALM method as best shown by the COIM method curves being above the other curves in these two plots. We also see for alternative $iv$, the most complex polynomial alternative included in the simulated data, that the Overall-$F$ method comes much closer in performance to the CALM and COIM methods. This coincides with our earlier statements that the Overall-$F$ method may have little power in detecting simple alternative patterns because it focuses its attention on more complex patterns. The ROC curves for the non-polynomial alternatives $v$ & $vi$ show a comparative performance that is similar to that found in the simpler alternatives of $i$ & $ii$.

In practice, a researcher will rank the pooled set of test statistics from all genes and reject in order starting with the most extreme and stopping when the rejection region coincides
with an acceptable number of significant genes and estimated number of false positives. For example, if we apply the CALM method to our simulated data and choose a $p$-value threshold of 0.005, then on average we reject 345 genes of which 329 are truly differentially expressed. This leads to an overall proportion of false positives of 4.6%. When the most extreme 345 test statistics are rejected for the COIM method, the false positive proportion is 8.4% and for the Overall-$F$ method the proportion is 28.1%.

The set of correctly rejected genes includes all types of alternative expressions. Though we generated the alternatives to be equally represented in each simulated data set, their representation in the set of rejected genes is not equal and depends on the type of alternative, the method applied, and the chosen significance threshold. In Table 2, we show the composition of the average rejection set for each method when the most extreme 345 test statistics are rejected in each simulation. The composition of the rejection region for the Overall-$F$ method coincides with earlier statements that this method focuses on detecting more complex alternatives. Similarly, alternative $iii$, which includes interaction, represents the largest portion of the COIM method rejection set, and the simplest polynomial alternative represents the largest portion of the CALM rejection set.

[Insert Table 2]

We computed the overall sensitivity for each method as the proportion of alternative genes that appeared in the aforementioned rejection region by considering the full set of 1200 alternative genes in every data set. On average, the sensitivity was 0.275, 0.264, and 0.206 for the CALM, COIM, and Overall-$F$ method, respectively. The related overall 1-specificity values were 0.0042, 0.0076, and 0.0255. We note that the results in Table 2
depend on the configuration of the simulated data, but by choosing the mean structures for
the first four alternatives to provide common Type I and Type II error rates, we attempted
to provide a reasonable comparison of the methods for these alternatives.

### 3.2.2 Model choice

The model choice candidate pool for each simulated data set included 13 polynomial regression models. We labeled the models from 1 to 13 to coincide with the ordered level of model complexity. We quantified complexity as the number of parameters required to fully describe a model’s mean structure. For models of equal complexity, we arbitrarily labeled the models using the appropriate consecutive numbers in the rank. Recalling that the null model is not included as a candidate, the simplest candidate, Model 1, included only a linear time effect. The most complex candidate, Model 13, included enough terms to describe an expression pattern represented by separate cubic polynomials for each treatment group.

The true models used to generate the mean structures of alternatives \( i \) through \( iv \) were present in the candidate pool. For this reason, we recorded the CALM method model choice at the genes where alternatives \( i \) through \( iv \) were simulated in order to estimate how often the CALM method chose the correct model. The models used to generate alternatives \( i \) through \( iv \) were Models 1, 5, 8, and 11, respectively, and Figure 2 shows the normalized histograms representing the frequency of model choice. In general, the CALM method chose the correct model a majority of the time for each of these alternatives. For Model 1, the CALM method chose the correct model more than 80% of the time. The histograms also show that when an incorrect model choice is made, the chosen model is usually close in complexity to the correct model.
Figure 2. Normalized histograms based on all 30 simulated data sets showing CALM method model choice frequencies at genes where alternatives i through iv were simulated. The model numbers representing the true models used to simulate alternatives i through iv are 1, 5, 8, and 11, respectively.

4 Example: Arabidopsis Experiment

We applied the CALM method to a data set generated from a microarray experiment designed to explore how different genetic lines of the Arabidopsis thaliana plant react to varying durations of ultra-violet light exposure. In this experiment, researchers harvested RNA from wildtype and two mutant lines of the plant at time points 0, 1, 4, 10, and 24 hours representing the duration of exposure. Fourteen plants were allocated to each line-time combination, but their RNA was pooled and hybridized to one array, leaving one expression measurement for each gene-line-time combination. Performance of this experiment required fifteen microarrays. Though there was no true replication, we were able to provide the researchers with a list of significantly differentially expressed genes and to estimate the expected number of false positives using our proposed method.

Each of the fifteen microarrays contained 8297 probe sets (or genes) and the expression at each gene was originally represented by the Affymetrix MAS 5.0 signal intensity (Affymetrix, 2001). To make measurements comparable across arrays, we logged and mean centered the MAS 5.0 values within each array. Finally, before applying our method to this data we
replaced each of the original five time points with its respective square root value in order to reduce the influence of the data point collected at 24 hours on the fitted polynomials.

The empirical distribution of $p$-values generated after applying the CALM method is shown in Figure 3A. Using the procedure presented by Storey and Tibshirani, we approximated the proportion of null genes to be 43% and computed a $q$-value for each gene. Figure 3B shows the relationship between the number of significant genes in the rejection region and the associated number of expected false positives to fall within that region.

[Insert Figure 3]

**Figure 3.** **A.** Distribution of 8297 $p$-values computed after applying the CALM method for locating differentially expressed genes to the *Arabidopsis* experiment. **B.** Plot showing the relationship between the expected number of false positives when a given number of genes is declared to be significantly differentially expressed estimated after applying the method proposed by Storey and Tibshirani (2003) to the distribution of $p$-values.

Choosing a $q$-value of 5.5% as the significance threshold provides us with a list of 500 significant genes. We chose three of these genes labeled as 4949, 8247, and 247 and plotted their observed expression patterns in Figure 4. The observed points are labeled with a 1, 2, or 3 coinciding with wildtype, mutant 1, and mutant 2 genetic lines, respectively. We then overlaid each gene’s fitted expression pattern based on the CALM method model choice. The fitted model for Gene 4949 contains only a linear time effect. The fitted model for Gene 8247 is represented by parallel lines for each genotype and the fitted model for Gene 247
includes interaction and is represented by non-parallel lines for each genotype.

[Insert Figure 4]

**Figure 4.** Observed expression patterns for Genes 4949, 8247, and 247. The labels of 1, 2, and 3 represent observations from the wildtype, mutant 1, and mutant 2 genetic lines, respectively. The overlaid black lines represent the CALM method fitted model for each gene.

## 5 Related Work

The two-stage approach we are proposing has some similarities to the two-stage approach proposed by Peddada et al. (2003) used to cluster and select genes according to their time-course expression. In their first step, the researcher chooses candidate profiles of interest represented by order-restricted means along the given time points. Thus, the shape of a candidate profile can be described as moving up, down, or staying the same for sequential observations in time. The choice of candidates appears to be biologically driven. In the second step, a best fitting profile is chosen for each gene from the pool of candidates, essentially clustering the genes into as many categories as there are candidate profiles. The authors then select genes only if their fit to the best candidate is stronger than the fit that would be seen when there was no correlation in the data. The strength of the fit is judged by comparing the observed fit statistic with a bootstrap null distribution of fit statistics.

Both our method and the method by Peddada et al. involve a model choice component followed by a hypothesis test based on the chosen model. The global null models for both
methods are similar in that the expected expression at a gene with a true null hypothesis is represented as a constant across all time points, but because our analysis includes a treatment effect, there is a definite difference in how these null models are interpreted. Accordingly, Peddada et al. focuses on the timecourse aspect by referring to the flat line expression as the *null profile model* and we focus on the concurrent movement of treatment groups as well as the time component by referring to the flat line expression as the *overall mean model*. We should point out, as Peddada et al. mentioned, that though our method is applied to a timecourse study, it could also be applied to a dose-response study, or to any experiment involving one categorical and one quantitative factor.

Park et al. (2003) present a method applied to multiple-treatment timecourse cDNA microarray data to identify genes with different expression profiles between treatment groups. As their first step, Park et al. dichotomizes the data into genes with and without significant treatment-by-time interaction using the relevant traditional $F$-test if the normality assumption is met or a permutation test if it is not. Genes in the first category are associated with the $p$-value testing for an interaction effect and genes in the second category are associated with the $p$-value testing for the main effect of treatment. The $p$-values are then adjusted for multiple comparisons and a $p$-value threshold is chosen to determine significance.

Similar to the overall $F$-test mentioned in Section 1, Park et al. focus on specific alternative patterns in the data by itemizing only two possible alternative hypotheses, whereas our method allows for numerous alternative hypotheses of interest. For example, our method has the ability to identify genes with concurrently upregulated or downregulated expression levels across all treatments, which is not an alternative pattern of interest in the method proposed by Park et al. Also, though they mentioned using their method for an unrepli-
ated experiment, to do so would require making the assumption of no treatment-by-time interaction, which is an assumption our method does not require.

6 Discussion

We have proposed a general method for identifying differentially expressed genes in a multiple-treatment timecourse microarray experiment where differential expression is defined as any expression profile that differs from the simple overall mean profile. Our method utilizes a permutation approach to first choose a best fitting model for each gene from a candidate pool of possible polynomial regression models and then generate a related permutation $p$-value to be used for testing the hypothesis of differential expression at each gene. Once the distribution of empirical $p$-values is generated, we suggest using the method proposed by Storey and Tibshirani (2003) to determine significance. This method uses the distribution of $p$-values to approximate the overall proportion of genes with a true null hypothesis. This subsequent estimate can then be used to help guide in the choice of a rejection region by providing an estimate for the number of expected false positives for any chosen significance threshold.

The candidate pool for model choice plays a large part in our method. By considering time as a continuous variable and including only models allowing for degrees of freedom for error in the candidate pool, we are able to apply our method to relevant unreplicated as well as replicated experiments. Also, after narrowing the candidate pool to a subset of the regression models, a researcher can apply our same proposed process, but focus more detection power on specific expression profiles. We exemplified this flexibility in our method.
by providing an example when a researcher would be most interested in detecting profiles that included interaction.

We chose the six alternatives included in our simulations to represent a variety of expression pattern complexities. As predicted, simulations showed our method to be more powerful in detecting relatively simple alternative patterns when compared to the method that uses the $F$-statistic from the overall $F$-test to detect differential expression. The results suggested that the advantage in our method diminishes as the alternative of interest becomes more complex. Further research may be conducted to determine the specific level of polynomial complexity at which the methods crossover in performance.

Undoubtedly, to gain power for detecting the simpler alternatives, our method must sacrifice some power for the more complex alternatives. To the researcher who is equally interested in a variety of expression patterns, this trade-off would seem worthwhile. Plus, as the number of treatments or time points in the design gets larger, the potential for net gain also increases. Regardless of which method is more powerful for a particular alternative, one advantage inherent in our method is that it offers a ‘best’ fitting model for every gene and as the simulations showed, its frequency of choosing the correct model for significant genes with a polynomial expression pattern appears to be high.

The computation time required to apply our method to a microarray data set depends on the experimental design and the chosen number of permutations. Using a 2.66 GHz dual processor machine, 2500 permutations, and the R environment (Ihaka and Gentleman, 1996), it took approximately 12 hours to complete a run on a single simulated data set. After switching to parallel R processing as described by Rossini, Tierney, and Li (2003) on a local cluster composed of four similar machines, a single run was completed in approxi-
mately 3 hours. We should also note the we were able to complete a run on a 3 GHz single processor machine using the open source programming language Python (van Rossum and Drake, 2002) in conjunction with Numerical Python (Ascher et al., 2001) in approximately 3.5 hours and this would be our preferred language for this method if restricted to a single processor. Related code is available from the first author upon request.

Acknowledgements

The authors would like to thank Howard Butler for his time and expertise in configuring a local cluster making parallel R computing possible.

References


Table 1

*Six alternative patterns included in the simulated data and the number of parameters required to fully describe their mean structures.*

<table>
<thead>
<tr>
<th>no.</th>
<th>Alternative pattern</th>
<th>Parameters in model</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>Linear in time</td>
<td>2</td>
</tr>
<tr>
<td>ii</td>
<td>Linear in time with treatment main effects</td>
<td>4</td>
</tr>
<tr>
<td>iii</td>
<td>Non-parallel lines for treatment groups</td>
<td>6</td>
</tr>
<tr>
<td>iv</td>
<td>Non-parallel quadratics for treatment groups</td>
<td>9</td>
</tr>
<tr>
<td>v</td>
<td>Concurrent jump at time midpoint</td>
<td>na</td>
</tr>
<tr>
<td>vi</td>
<td>Non-concurrent logistic growth curves</td>
<td>na</td>
</tr>
</tbody>
</table>
Table 2

Composition of the average rejection set for each method when the most extreme test statistics are rejected (provided as a count and as a percentage of correct rejections).

<table>
<thead>
<tr>
<th>Alternative</th>
<th>CALM Count</th>
<th>%</th>
<th>COIM Count</th>
<th>%</th>
<th>Overall-F Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>81</td>
<td>24.6</td>
<td>30</td>
<td>9.4</td>
<td>19</td>
<td>7.8</td>
</tr>
<tr>
<td>ii</td>
<td>76</td>
<td>23.0</td>
<td>68</td>
<td>21.6</td>
<td>31</td>
<td>12.6</td>
</tr>
<tr>
<td>iii</td>
<td>50</td>
<td>15.3</td>
<td>96</td>
<td>30.3</td>
<td>50</td>
<td>20.2</td>
</tr>
<tr>
<td>iv</td>
<td>23</td>
<td>7.1</td>
<td>59</td>
<td>18.8</td>
<td>105</td>
<td>42.1</td>
</tr>
<tr>
<td>v</td>
<td>37</td>
<td>11.3</td>
<td>10</td>
<td>3.2</td>
<td>8</td>
<td>3.1</td>
</tr>
<tr>
<td>vi</td>
<td>62</td>
<td>18.7</td>
<td>53</td>
<td>16.7</td>
<td>35</td>
<td>14.2</td>
</tr>
<tr>
<td>Total</td>
<td>329</td>
<td>100.0</td>
<td>316</td>
<td>100.0</td>
<td>248</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Figure 1.

ROC curves for each method at each of the six alternatives $i$ through $vi$ included in the simulated data based on the average of 30 simulated data sets. The solid line, dotted line, and dashed line represent the CALM method, COIM method, and Overall-$F$ method, respectively.
Figure 2.

Normalized histograms based on all 30 simulated data sets showing CALM method model choice frequencies at genes where alternatives $i$ through $iv$ were simulated. The model numbers representing the true models used to simulate alternatives $i$ through $iv$ are 1, 5, 8, and 11, respectively.
Figure 3.

A. Distribution of 8297 $p$-values computed after applying the CALM method for locating differentially expressed genes to the *Arabidopsis* experiment. B. Plot showing the relationship between the expected number of false positives when a given number of genes is declared to be significantly differentially expressed estimated after applying the method proposed by Storey and Tibshirani (2003) to the distribution of $p$-values.
**Figure 4.**

Observed expression patterns for Genes 4949, 8247, and 247. The labels of 1, 2, and 3 represent observations from the wildtype, mutant 1, and mutant 2 genetic lines, respectively. The overlaid black lines represent the CALM method fitted model for each gene.