BIOINFORMATICS

Vol. 19 no. 7 2003, pages 834–841
DOI: 10.1093/bioinformatics/btg093

Gene selection and clustering for time-course and dose–response microarray experiments using order-restricted inference

Shyamal D. Peddada[1,*], Edward K. Lobenhofer[2], Leping Li[1], Cynthia A. Afshari[2,3,†], Clarice R. Weinberg[1] and David M. Umbach[1]

[1]Biostatistics Branch, [2]Gene Regulation Group, Laboratory of Molecular Carcinogenesis and [3]NIEHS Microarray Center, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA

Received on August 8, 2002; revised on November 14, 2002; accepted on November 29, 2002

ABSTRACT

We propose an algorithm for selecting and clustering genes according to their time-course or dose–response profiles using gene expression data. The proposed algorithm is based on the order-restricted inference methodology developed in statistics. We describe the methodology for time-course experiments although it is applicable to any ordered set of treatments. Candidate temporal profiles are defined in terms of inequalities among mean expression levels at the time points. The proposed algorithm selects genes when they meet a bootstrap-based criterion for statistical significance and assigns each selected gene to the best fitting candidate profile. We illustrate the methodology using data from a cDNA microarray experiment in which a breast cancer cell line was stimulated with estrogen for different time intervals. In this example, our method was able to identify several biologically interesting genes that previous analyses failed to reveal.

Contact: peddada@embryo.niehs.nih.gov

INTRODUCTION

A number of methods have been proposed for selecting genes that exhibit interesting changes in expression between classes of samples. Depending on the data available, any of these methods can be employed to select genes that are differentially expressed across time points. Examples of these methods include the standard two-sample t-test and its modifications (Golub *et al.*, 1999; Long *et al.*, 2001; Tusher *et al.*, 2001) and a confidence interval method (Chen *et al.*, 1997). A different approach to the selection of a subset of discriminative genes is the multivariate genetic algorithm/k-nearest neighbors methodology (Li *et al.*, 2001a,b).

An important application of microarray technology is to study patterns of gene expression across a series of time points or of doses levels. The premise is that genes sharing similar expression profiles might be functionally related or co-regulated. Therefore, microarray data may provide insight into gene–gene interactions, gene function and pathway identification. In toxicogenomics, these studies can also provide information about the dynamic responses of cells (tissues) to chemical insults (Hamadeh *et al.*, 2001). We focus on time-course studies although our methodology is applicable to dose-response studies as well. None of the previously mentioned methods, however, take advantage of the ordering in a time-course study. In contrast to those methods, explicit use of the temporal ordering should allow more sensitive detection of genes that display a consistent pattern over time.

Some authors have developed correlation-based methods for clustering genes with similar temporal profiles (Chu *et al.*, 1998; Heyer *et al.*, 1999). Chu *et al.* (1998) applied their methodology to select genes from a yeast cell line into seven temporal patterns of expression. Their approach pre-identifies a few candidate temporal profiles along with a sample of three to eight genes per profile. Using these template genes, they estimate the mean expression at each time point for each profile. Thus, each candidate profile is defined by an estimated time-course curve. Each remaining gene is then either assigned to one of the candidate profiles or not assigned into any, depending upon the magnitude of the correlation coefficient between the gene's experimentally determined profile and each of the candidate profiles. Heyer *et al.* (1999), employing a jack-knifed correlation coefficient, also proposed a procedure for clustering genes from time-course

*To whom correspondence should be addresed.
† Present address: Amgen Inc. Thousand Oaks, CA 91320, USA.

Published by Oxford University Press

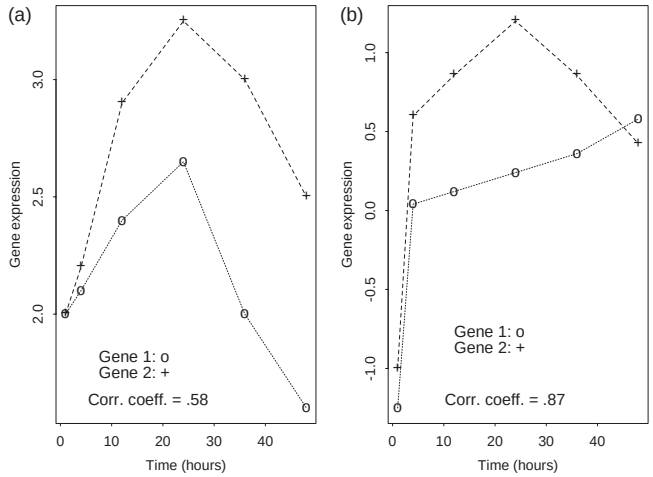

Fig. 1. (**a**) Two genes with similar profiles (maxima at 24 hours) that may not be clustered together by correlation-based methods. (**b**) Two genes with different profiles (monotone versus up–down) that are likely to be clustered together by correlation-based methods.

experiments. Although their basic procedure did not require candidate profiles, they describe a modification where the clustering algorithm is seeded with candidate profiles.

Correlation-based procedures using candidate profiles require the scientist to specify expression levels defining each profile in advance. This requirement means that researchers must estimate each candidate profile using a small sample of handpicked genes. More importantly, the clustering that results depends upon the genes initially chosen as templates; thus, important genes may be missed. Furthermore, the sample size for each correlation coefficient is the number of time points, not the number of actual observations. The correlation coefficient may not be a reliable measure of association when an experiment has few time points. Moreover, a large correlation coefficient does not necessarily indicate two similarly shaped profiles, nor does a small correlation coefficient necessarily indicate differently shaped profiles. Figure 1a and b each shows hypothetical mean expression levels for two genes at six time points. The two genes in Figure 1a arguably display similar patterns, in that both attain a peak value at the 4th time point. Yet their correlation coefficient is only 0.58, suggesting that they might not be grouped together by correlation-based methods. On the other hand, the two genes in Figure 1b display apparently different patterns over time. One increases monotonically whereas the other has a peak at the 4th time point, yet they have a high correlation coefficient of 0.87 and would likely be clustered together by a correlation-based approach. Thus, correlation-based methods may either miss some important genes or cluster genes with different profiles.

Herein, we propose an alternative methodology to select and cluster genes using the ideas of order-restricted inference, where estimation makes use of known inequalities among parameters. The first step is to define potential candidate profiles of interest and to express them in terms of inequalities between the expected gene expression levels at various time points. For a given candidate profile, we estimate the mean expression level of each gene using the procedure developed in Hwang and Peddada (1994). The best fitting profile for a given gene is then selected using the goodness-of-fit criterion and the bootstrap test procedure developed in Peddada *et al.* (2001). A pair of genes g_1 and g_2 fall into the same cluster if all the inequalities between the expected expression levels at various time points are the same, that is, if they follow the same temporal profile. In this sense, the genes of Figure 1a are similar, while those of Figure 1b are not. Our procedure is less restrictive than those that define profiles via pre-specified expression levels because only the general shape of the profile is needed.

METHODOLOGY

Suppose a time-course experiment includes T time points denoted by $1, 2, \ldots, T$, and at each time point there are M arrays, each with G genes. Let Y_{igt} denote the ith expression measurement taken on gene g at time point t. Let $\bar{Y}_{gt}$ denote the sample mean of gene g at time point t and let $\bar{Y}_g = (\bar{Y}_{g1}, \bar{Y}_{g2}, \ldots, \bar{Y}_{gT})'$. The unknown true mean expression level of gene g at time point t is $E(\bar{Y}_{gt}) = \mu_{gt}$. Inequalities between the components of $\mu_g = (\mu_{g1}, \mu_{g2}, \ldots, \mu_{gT})'$ define the true profile for gene g. Our procedure seeks to match a gene's true profile, estimated from the observed data, to one of a specified set of candidate profiles.

Examples of inequality profiles are given below. For simplicity, we often drop the subscript g.

Null profile: $C_0 = \{\mu \in R^T : \mu_1 = \mu_2 = \cdots = \mu_T\}$.

Monotone increasing profile (simple order):

$$C = \{\mu \in R^{\mathrm{T}} : \mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_T\} \qquad (1)$$

(with at least one strict inequality). One may similarly define a *monotone decreasing profile* by replacing $\leqslant$ by $\geqslant$ in Equation (1).

Up–down profile with maximum at i (umbrella order):

$$C = \{\mu \in R^{\mathrm{T}} : \mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_i \geqslant \mu_{i+1} \geqslant \cdots \geqslant \mu_T\} \qquad (2)$$

(with at least one strict inequality among $\mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_i$ and one among $\mu_i \geqslant \mu_{i+1} \geqslant \cdots \geqslant \mu_T$).

Genes satisfying this profile have mean expression values non-decreasing in time up to time point i and

non–increasing thereafter. One may similarly define a *down–up profile with minimum at i*.

Cyclical profile with minima at 1, j, and T and maxima at i and k:

$$C = \{\mu \in R^{\mathrm{T}} : \mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_i \geqslant \mu_{i+1} \geqslant \cdots \geqslant \mu_j \leqslant \mu_{j+1} \leqslant \cdots \leqslant \mu_k \geqslant \mu_{k+1} \geqslant \cdots \geqslant \mu_T\} \quad (3)$$

(with at least one strict inequality among each monotone sub-profile). Cyclical profiles may be important in long time-course experiments where the mean expression value could oscillate.

Incomplete inequality profiles:

$$C = \{\mu \in R^{\mathrm{T}} : \mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_i, \mu_{i+1} \geqslant \cdots \geqslant \mu_j, \mu_{j+1} \leqslant \cdots \leqslant \mu_k, \mu_{k+1} \geqslant \cdots \geqslant \mu_T\} \quad (4)$$

(with at least one strict inequality among each monotone sub-profile). Profiles like Equation (4) are useful when the investigator is unable to specify inequalities between certain means.

For compactness, we drop $\mu \in R^{\mathrm{T}}$ and the phrase 'with a strict inequality'.

DEFINITION 1. Two parameters in a given profile are said to be *linked* if the inequality between them is specified *a priori*.

DEFINITION 2. For a given profile, a parameter is said to be *nodal* if it is linked with every other parameter in the profile.

For example, μ_i is the only nodal parameter in Equation (2) while there are no nodal parameters in Equation (3).

DEFINITION 3. Define the ℓ_∞ norm of an estimated profile as the maximum difference between the estimates of two linked parameters.

Other norms could replace ℓ_∞. Our choice is motivated by its connection to well-known procedures for order-restricted inference. For example, Williams' test for trend in normal means (Williams, 1977) and Dunnett's multiple comparison test procedure (Dunnett, 1955) are based on ℓ_∞ norm. In the case of profile Equation (2), $\ell_\infty = \max\{\hat{\mu}_i - \hat{\mu}_1, \hat{\mu}_i - \hat{\mu}_T\}$, where $\hat{\mu}_j$ is an estimate of μ_j, $j = 1, 2, \ldots, T$.

DEFINITION 4. An inequality *sub-profile* C_i within a profile C is described by the inequalities between the components of the sub-vector $\mu_i = (\mu_{i_1}, \mu_{i_2}, \ldots, \mu_{i_s})$, where $\{i_1, i_2, \ldots, i_s\} \subseteq \{1, 2, \ldots, T\}$.

The proposed algorithm

STEP 1. Pre-specify a collection of candidate profiles. Denote these profiles by $C_1, C_2, \ldots, C_p$.

EXAMPLE 2.1. Suppose an experiment consists of four time points at 1, 2, 3 and 4 hours, and we are interested in identifying genes belonging to either of the following profiles: $C_1 = \{\mu \in R^{\mathrm{T}} : \mu_1 \leqslant \mu_2 \geqslant \mu_3 \geqslant \mu_4\}$, $C_2 = \{\mu \in R^{\mathrm{T}} : \mu_1 \leqslant \mu_2 \leqslant \mu_3 \geqslant \mu_4\}$.

For each gene g, $g = 1, 2, \ldots, G$, perform the following steps.

STEP 2. Obtain the estimates of $\mu_{g1}, \mu_{g2}, \ldots, \mu_{gT}$ under each of the candidate profiles $C_1, C_2, \ldots, C_p$ using Hwang and Peddada (1994). See the Appendix for details.

EXAMPLE 2.1 (CONTINUED). Suppose the sample mean expression levels of a gene at the four time points are 0.2, 0.4, 0.8, and 0.5, respectively. Assume the sample sizes are equal for all time points.

Estimation under C_1 Since μ_2 is the only nodal parameter in C_1, we first estimate μ_2. Maintaining all the known inequalities in C_1 and assigning arbitrary inequalities where they are unknown, we take $\mu_4 \leqslant \mu_1 \leqslant \mu_3 \leqslant \mu_2$. Using formula (A2) in the appendix, we obtain $\hat{\mu}_2 = 0.6$. Now estimate μ_3 and μ_4, nodal parameters in the sub-profile $\mu_4 \leqslant \mu_3 \leqslant \mu_2$, and μ_1, a nodal parameter in the sub-profile $\mu_1 \leqslant \mu_2$. Using Equation (A3) and (A4) in the appendix, we obtain $\hat{\mu}_1 = 0.2$, $\hat{\mu}_3 = 0.6$, $\hat{\mu}_4 = 0.5$.

Estimation under C_2 In this case, $\hat{\mu}_1 = 0.2, \hat{\mu}_2 = 0.4, \hat{\mu}_3 = 0.8, \hat{\mu}_4 = 0.5$.

STEP 3. For each C_i, $i = 1, 2, \ldots p$, compute $\ell_\infty^{g(i)}$. Let r be such that $\ell_\infty^{g(r)} = \max_i \ell_\infty^{g(i)}$.

EXAMPLE 2.1 (CONTINUED). Here $\ell_\infty^{g(1)} = 0.4$, $\ell_\infty^{g(2)} = 0.6$, hence $\ell_\infty^{g(r)} = 0.6$ and $r = 2$.

STEP 4 (BOOTSTRAP NULL DISTRIBUTION). Assuming that the true means and variances are the same at every time point, draw N bootstrap samples. Each bootstrap sample is obtained as follows. Combine the actual observations from all the time points into a vector of length MT and draw T simple random samples with replacement, each of size M. Repeat Steps 2 and 3 for each bootstrap sample. This results in a bootstrap distribution for $\max_i \ell_\infty^{g(i)}$, which is used for testing

$$H_0 : \mu \in C_0, H_a : \mu \in \bigcup_{i=1}^{p} C_i \quad (5)$$

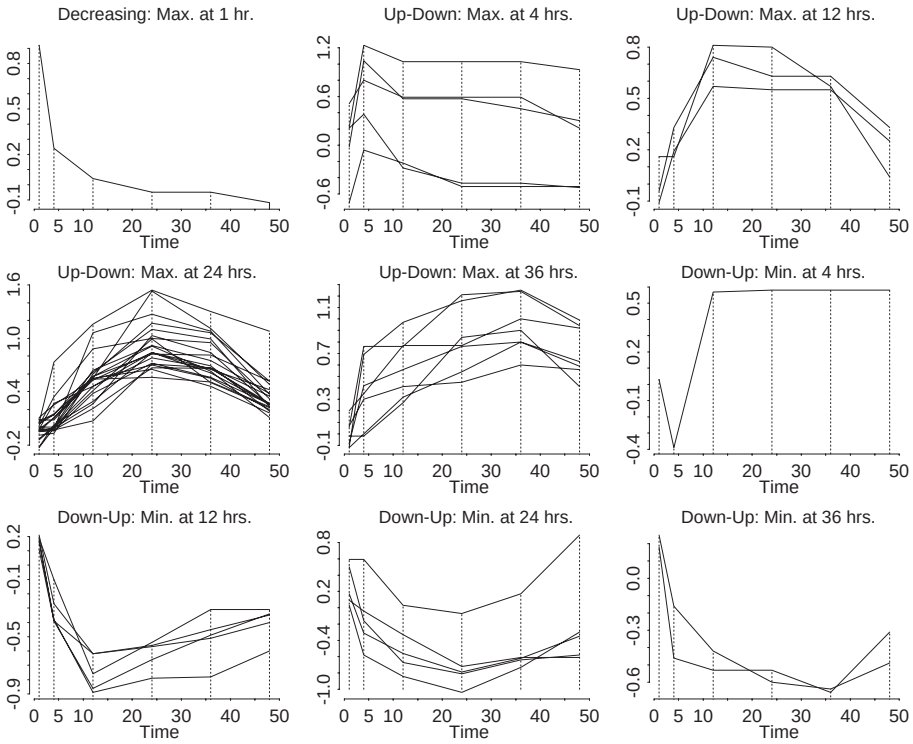

Fig. 2. Estimated profiles of the selected top 50 genes. Curves represent order-restricted estimates of mean log expression ratios. Vertical lines correspond to the six time points.

STEP 5. Assign gene g to profile C_r if $\ell_\infty^{g(r)} \geqslant z_\alpha^*$, where z_α^* is the upper αth percentile of the bootstrap distribution derived in Step 4. If $\ell_\infty^{g(r)} \leqslant z_\alpha^*$ or if two profiles are tied then do not classify g into any of the p profiles.

STEP 6. Repeat Steps 2–5 with every gene.

STEP 7 (OPTIONAL). Some genes selected by the above process may have small mean expression levels at each time point. Some investigators may want to restrict attention to those genes that have large expression levels at one time point at least. If so, we suggest the following filtering process. If the data are centered, then for each gene g selected after Step 6, let $t_g = \sum_{i=1}^{T} \hat{\mu}_i^2$; alternatively, let

$$t_g = \sum_{i=1}^{T} (\hat{\mu}_i - \bar{\hat{\mu}})^2, \quad \text{where} \quad \bar{\hat{\mu}} = \frac{1}{T} \sum_{i=1}^{T} \hat{\mu}_i .$$

Large values of t_g indicate that the mean expression of gene g is high for at least one time point. Arrange the genes in descending order of t_g and retain the top R genes.

APPLICATION TO BREAST CANCER CELL LINE DATA

We illustrate the proposed methodology using log-transformed relative expression data from Lobenhofer *et al.* (2002). In that study, the MCF-7 breast cancer cell line was treated with 17β-estradiol or ethanol (vehicle control). Samples were harvested at 1, 4, 12, 24, 36 and 48 hours after treatment. At each time point $M = 8$ hybridizations were performed. Each array consisted of $G = 1900$ genes. For each gene, we assumed that the variance of the log relative expression was homoscedastic over time. For each gene g, we tested Equation (5) where the alternative hypothesis is the union of the following 10 profiles: monotone decreasing, C_1; monotone increasing, C_2; four up–down profiles with maxima at 4, 12, 24, 36 hours, C_3–C_6, respectively; and 4 down–up profiles with minima at 4, 12, 24, and 36 hours, C_7–C_{10}, respectively.

Using Steps 1–6 with $N = 50\,000$, we obtained 124 genes with a p-value $\leqslant$0.0025. Of these, 10 were clustered into C_1, 14 into C_2, four into C_3, 31 into C_4, 12 into C_5, one into C_7, nine into C_8, 34 into C_9 and nine into C_{10}. Applying Step 7 we selected the top 50 genes among these 124. These 50 genes display nine of the 10 candidate profiles (Table 1, Fig. 2).

Table 1. Genes classified according to inequality profile

Clone ID	Gene name	Functional category	Previously identified
Decreasing with maximum at 1 hour			
417226	v-myc viral oncogene homolog	Transcription/Chromatin Structure	Yes
Up–Down with maximum at 4 hours			
110022	Cyclin D1	Cell Cycle	Yes
428733	Protein kinase C, delta	Cellular Signaling	Yes
362059	Laminin, alpha 3, kalinin, epilegrin	Extracellular Matrix/Cell Structure	Yes
417503	EST	Unknown	Yes
248613	v-myb viral oncogene homolog	Transcription/Chromatin Structure	No
Up–Down with maximum at 12 hours			
563187	CDC6	Cell Cycle	Yes
321207	Polymerase (DNA directed), epsilon	DNA Replication/Repair	Yes
196676	Replication factor C (activator 1) 4	DNA Replication/Repair	No
Up–Down with maximum at 24 hours			
129140	MAD2L1	Cell Cycle	Yes
248008	Deoxythymidylate kinase	Cell Cycle	Yes
489092	Deoxythymidylate kinase	Cell Cycle	No
285427	CSE1L	Cell Cycle	Yes
359119	CDC28 protein kinase 2	Cell Cycle	Yes
415639	Serine/threonine kinase 15	Cell Cycle	Yes
488059	Tubulin, gamma 1	Cell Cycle	Yes
563809	CDC20	Cell Cycle	Yes
293274	Cyclin-dependent kinase inhibitor 3	Cell Cycle	No
49950	Flap structure-specific endonuclease 1	DNA Replication/Repair	Yes
346838	Minichromosome maintenance deficien 3	DNA Replication/Repair	Yes
359465	Dihydrofolate reductase	DNA Replication/Repair	Yes
487757	Ligase I, DNA, ATP-dependent	DNA Replication/Repair	Yes
49940	Replication factor C (activator 1) 5	DNA Replication/Repair	No
52713	Vitronectin	Extracellular Matrix/Cell Structure	Yes
339075	Karyopherin alpha 2	Protein Degradation/Synthesis/Targeting	Yes
136609	v-myb homolog-like 1	Transcription/Chromatin Structure	Yes
198205	v-myb homolog-like 2	Transcription/Chromatin Structure	Yes
229509	coagulation factor V	Miscellaneous	No
200573	EST	Unknown	Yes
366842	EST	Unknown	No
Up–Down with maximum at 36 hours			
264117	Cathepsin D	Cell Cycle	Yes
150163	Neuropeptide Y receptor Y1	Cellular Signaling	Yes
238545	ADP-ribosylation factor-like 3	Cellular Signaling	Yes
242182	Protein kinase inhibitor beta	Cellular Signaling	Yes
509614	High-mobility group protein 1	Transcription/Chromatin Structure	Yes
510595	Lactate dehydrogenase A	Miscellaneous	Yes
470480	Autocrine motility factor receptor	Miscellaneous	No
Down-Up with minimum at 4 hours			
487407	Insulin induced gene 1	Miscellaneous	Yes
Down–Up with minimum at 12 hours			
361381	Myeloid cell leukemia sequence 1	Apoptosis	Yes
145093	Myeloid cell leukemia sequence 1	Apoptosis	No
485875	EFEMP1	Extracellular Matrix/Cell Structure	Yes
34821	CHRNA 4	Miscellaneous	Yes
Down–Up with minimum at 24 hours			
359191	Protein kinase H11	Cellular Signaling	Yes
180789	Low density lipoprotein-related protein 1	Protein Degradation/Synthesis/Targeting	Yes

Table 1. *Continued.*

Clone ID	Gene name	Functional category	Previously identified
162479	E74-like factor 3	Transcription/Chromatin Structure	Yes
430235	H2B histone family, member Q	Transcription/Chromatin Structure	Yes
545242	STAT 1	Transcription/Chromatin Structure	Yes
268652	p21/CIP 1	Cell Cycle	No
Down–Up with minimum at 36 hours			
29682	Protein kinase C binding protein 1	Cellular Signaling	Yes
365147	v-erb-b2 homolog 2	Cellular Signaling	No

The confidence-interval approach of Chen *et al.* (1997) identified 105 genes that demonstrated estrogen-responsive expression (Lobenhofer *et al.*, 2002). Of these 105, 39 were also among our top 50. Most of the 39 genes selected in common are involved in cell cycle progression and DNA replication (Lobenhofer *et al.*, 2002), reflecting the known sensitivity of MCF-7 cells to estrogen.

Most of our 11 newly identified genes also display typical phenotypes for estrogen-treated MCF-7 cells. For example, one initial step in DNA replication is the binding of a complex of proteins (known as replication factor C) to DNA in order to recruit other proteins necessary for DNA synthesis. The confidence-interval approach identified one subunit (replication factor C 3) as being regulated by estrogen. Our order-restricted-inference approach identified an additional two components of the complex (replication factors C 4 and C 5) as having increased levels of expression at time points when the estrogen-stimulated cells are undergoing DNA synthesis. Another interesting observation was the decreased expression of cyclin-dependent kinase inhibitor 1A (p21 and Cip1), as shown previously by Prall *et al.* (2001). This inhibitory gene not only functions in the cell cycle at the transition from the G1 into the S phase (during which genome replication occurs) but also in the process of DNA synthesis. Therefore, the fact that estrogen induces MCF-7 cells to divide supports the finding that a gene that inhibits this process would be repressed. Finally, several genes were represented by two different spots (clones) on the microarray chips. Using the confidence-interval approach, deoxythymidylate kinase (248008) and myeloid cell leukemia sequence 1 (361381) were seen to be regulated by estrogen. Interestingly, the order-restricted-inference approach not only identified these genes as exhibiting altered expression in the presence of estrogen but also identified them based on two different spots that represent the same genes (Clone IDs 489092 and 145093; Table 1). These findings illustrate that our methodology can identify genes whose estrogen responses are biologically interpretable.

A simulation experiment

We investigated the false positive rate of our procedure using a small simulation study. To generate unpatterned null data, we created 48 new observations for each gene by randomly assigning the 48 original observations (with replacement), eight to each of the six time points. By this device, we generated 1900 genes whose true underlying profiles lack any systematic pattern. Our simulations suggest that our methodology provides fairly accurate type I error rates and tends to be conservative for smaller levels of significance. For example, corresponding to a nominal level of 0.0025, our estimated Type I error was 0.0005; and, for a nominal level of 0.05, our estimated Type I error was 0.049.

DISCUSSION

In studies where experimental conditions have an inherent ordering, making use of ordering information can improve inference. In microarray experiments, the ability to exploit ordering information may be especially valuable because genes whose expression levels change in concert through time may be components of the same cellular process or may share regulatory elements. Yet, virtually none of the commonly used methods for analysis of microarray data take account of time ordering. Those researchers who have recognized the importance of time-course information (Chu *et al.*, 1998; Heyer *et al.*, 1999) developed procedures based on correlation coefficients, an approach fundamentally different from ours. We have proposed an algorithm based on the statistical theory of order-restricted inference that makes explicit use of ordering information when selecting differentially expressed genes. Our approach selects genes whose expression levels through time are both significantly different from the null profile and similar to one of a set of pre-identified candidate profiles. Consequently, selected genes are naturally clustered into classes with similarly shaped profiles.

Our methodology is general and enjoys several desirable properties. First, the estimated mean expression levels, subject to an inequality profile, satisfy certain optimality

properties discussed in Hwang and Peddada (1994). In particular, the estimator *universally dominates* the unrestricted maximum likelihood estimator. Secondly, genes are selected into clusters based in part on a statistical significance criterion. Groupings obtained using unsupervised methods such as cluster analytic algorithms cannot make claims about Type I error rates. A related and important feature of our procedure is that it can select genes with subtle but reproducible expression changes over time and, hence, uncover some genes that may not be detectable by other approaches. Our example illustrated this feature with respect to the approach of Chen *et al.* (1997).

Both our procedure and correlation-based procedures require that investigators pre-specify a set of candidate profiles. What exactly is required of the candidate profiles, however, differs sharply between the two approaches. With our procedure, one need only describe the shapes of profiles in terms of mathematical inequalities; whereas, with the correlation-based procedures, one must specify numerical values at each time point for each candidate profile. Since exact values at time points are rarely known *a priori*, correlation-coefficient-based procedures often use averages from selected small subsets of genes to establish candidate profiles. Those genes that establish the profiles are essentially exempted from the analysis, and they may be the only genes representing their profiles. Our candidate profiles are specified without reference to data from the study, and a candidate profile may turn out to be represented by no genes. In fact, no genes in the top 50 of our example followed the monotone non-decreasing profile. Thus, our methodology is much less restrictive than the correlation-coefficient-based alternative.

Kerr and Churchill (2001) have advocated investigating the reliability of clustering results from microarray studies. Although we have not formally examined the reliability of our clustering results in that sense, one can conceive of embedding our procedure into a general bootstrapping framework similar in spirit to their approach.

Investigators might be interested in distinguishing more subtle patterns than considered here. For example, Chu *et al.* (1998) display two candidate profiles (Early I and Early II in their Figure 4(b) that rise to a maximum at 7 hours and then decrease. Thus, both are up–down profiles with a maximum at 7 hours and would not be distinguished by the candidate profiles that we have described. These two profiles differ, however, in that one rises rapidly after the first time point then more slowly to the peak whereas the other exhibits a rapid rise after the second time point. Our approach could be adapted to distinguish such sub-profiles by imposing order restrictions on suitable differences among mean expression levels.

Although the procedure that we described is designed for genes with a constant variance through time, it can be generalized to handle situations when the variances change or are subject to order restrictions themselves. In such situations, the estimation of mean gene expression outlined in this paper may be modified along the lines of Shi (1994). The required modifications to the bootstrap described in Step 4 remain a subject for future investigation.

In conclusion, we believe that methods of analysis that exploit the ordering of treatments to improve estimation will become increasingly valuable for time-course and dose–response microarray experiments. Our approach based on order-restricted inference should improve gene selection and clustering whenever treatments are inherently ordered.

ACKNOWLEDGEMENTS

The authors thank Drs D. Dunson, K. Kerr, F. Parham, N. Walker and R. Wolfinger for their careful reading of this manuscript and for their useful comments that improved the presentation.

REFERENCES

Chen,Y., Dougherty,E. and Bittner,M. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics*, **2**, 364–374.

Chu,S., DeRisi,J., Eisen,M., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.

Dunnett,C. (1955) A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Statist. Assoc.*, **50**, 1096–1121.

Golub,T., Slonim,D., Tamayo,P., Huard,C., Caasenbeek,J., Coller,H., Loh,M., Downing,J., Caliguiri,M., Bloomfield,C. and Lander,E. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hamadeh,H., Bushel,P., Paules,R. and Afshari,C. (2001) Discovery in toxicology: mediation by gene expression array technology. *J. Biochem. Molec. Tox.*, **15**, 231–242.

Heyer,L.J., Kruglyak,S. and Yooseph,S. (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.*, **9**, 1106–1115.

Hwang,J. and Peddada,S. (1994) Confidence interval estimation subject to order restrictions. *Ann. Statist.*, **22**, 67–93.

Kerr,M. and Churchill,G. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl Acad. Sci. USA*, **98**, 8961–8965.

Li,L., Weinberg,C., Darden,T. and Pedersen,L. (2001a) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, **17**, 1131–1142.

Li,L., Darden,T., Weinberg,C., Levine,A. and Pedersen,L. (2001b) Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. *Comb. Chem. High Throughput Screening*, **4**, 727–739.

Lobenhofer,E., Bennett,L., Cable,P,, Li,L., Bushel,P, and Afshari,C, (2002) Regulation of DNA replication fork genes by 17beta-estradiol. *Molec. Endocrin.*, **16**, 1215–1229.

Long,A., Mangalam,H., Chan,B., Tolleri,L., Hatfield,G. and Baldi,P. (2001) Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in *Escherichia coli K12*. *J. Biol. Chem.*, **276**, 19 937–19 944.

Peddada,S., Prescott,K. and Conaway,M. (2001) Tests for order restrictions in binary data. *Biometrics*, **57**, 1219–1227.

Prall,O., Carroll,J. and Sutherland,R. (2001) A low abundance pool of nascent p21WAF1/Cip1 is targeted by estrogen to activate cyclin E*Cdk2. *J. Biol. Chem.*, **276**, 45 433–45 442.

Shi,N. (1994) Maximum likelihood estimation of means and variances from normal opulations under simultaneous order estrictions. *J. Mult. Anal.*, **50**, 282–293.

Tusher,V., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Williams,D. (1977) Some inference procedures for monotonically ordered normal means. *Biometrika*, **64**, 9–14.

APPENDIX: ESTIMATION OF PARAMETERS (HWANG AND PEDDADA, 1994)

There are two types of profiles, those with at least one nodal parameter and those with no nodal parameters. We first describe estimation for the former.

(A) PROFILES WITH AT LEAST ONE NODAL PARAMETER

For a gene g at time t, suppose $\bar{Y}_t$ is the sample mean based on n_t observations. We assume that $\mathrm{Var}(\bar{Y}_t) = \frac{\sigma^2}{n_t}$. Repeat Steps A1–A4 described below until all parameters are estimated.

Estimation of nodal parameters in a given inequality profile

STEP A1. Suppose μ_t is a nodal parameter in the profile. Maintaining all the known inequalities and assigning arbitrary inequalities among those parameters where the inequalities are unknown, one obtains a non-decreasing order of the form Equation (1). For $i = 1, 2, \ldots, T$, let the ordered true means be denoted by $\mu_{(i)}$ where, $\mu_{(1)} \leqslant \mu_{(2)} \leqslant \cdots \leqslant \mu_{(T)}$, and the corresponding sample means and sample sizes be denoted by $\bar{Y}_{(i)}$, and $n_{(i)}$, respectively. Thus, for some s, $\mu_t \equiv \mu_{(s)}$.

EXAMPLE A1. Consider a profile with four parameters such that $\mu_1 \leqslant \mu_2 \geqslant \mu_3 \geqslant \mu_4$. Here the only nodal parameter is μ_2. Inequalities between μ_1, μ_3 and between μ_1, μ_4 are unknown. To estimate μ_2, we may arrange the four parameters as $\mu_4 \leqslant \mu_1 \leqslant \mu_3 \leqslant \mu_2$. Thus $\mu_4 \equiv \mu_{(1)}, \mu_1 \equiv \mu_{(2)}, \mu_3 \equiv \mu_{(3)}, \mu_2 \equiv \mu_{(4)}$.

STEP A2. Estimate the nodal parameter $\mu_t \equiv \mu_{(s)}$ using the following formula:

$$\hat{\mu}_t \equiv \hat{\mu}_{(s)} = \min_{r \geqslant s} \max_{q \leqslant s} \frac{\sum_{k=q}^{r} n_{(k)} \bar{Y}_{(k)}}{\sum_{k=q}^{r} n_{(k)}} \tag{A1}$$

In the example, the estimate of $\mu_2 \equiv \mu_{(4)}$ is

$$\hat{\mu}_2 \equiv \hat{\mu}_{(4)} = \max_{q \leqslant 4} \frac{\sum_{k=q}^{4} n_{(k)} \bar{Y}_{(k)}}{\sum_{k=q}^{4} n_{(k)}} \tag{A2}$$

STEP A3. Once a parameter is estimated, in all future calculations replace its sample mean $\bar{Y}$ by its estimated value $\hat{\mu}$ from Step A2, and its sample size n by B, where $B \to \infty$.

Estimation of non-nodal parameters

STEP A4. To estimate a non-nodal parameter μ_t, identify the largest sub-profile having μ_t as a nodal parameter. Using the data corresponding to the sub-profile, estimate μ_t by applying Steps A1–A3.

EXAMPLE A1 (CONTINUED). The largest sub-profile in which μ_3 is nodal is also the largest in which μ_4 is nodal: $\mu_4 \leqslant \mu_3 \leqslant \mu_2$. Hence μ_3 and μ_4 are estimated using formulae derived from (A1):

$$\hat{\mu}_3 = \min\left\{\max\left\{\bar{Y}_3, \frac{n_3\bar{Y}_3 + n_4\bar{Y}_4}{n_3 + n_4}\right\}, \hat{\mu}_2\right\},$$

$$\hat{\mu}_4 = \min\left\{\bar{Y}_4, \frac{n_3\bar{Y}_3 + n_4\bar{Y}_4}{n_3 + n_4}, \hat{\mu}_2\right\} \tag{A3}$$

Note that μ_1 is nodal in the sub-profile $\mu_1 \leqslant \mu_2$. Hence, from (A1) we deduce:

$$\hat{\mu}_1 = \min\{\bar{Y}_1, \hat{\mu}_2\} \tag{A4}$$

(B) PROFILES WITH NO NODAL PARAMETERS

STEP B1. Identify the largest sub-profile with at least one nodal parameter. Estimate all parameters of the sub-profile using Steps A1–A4.

STEP B2. Identify the next largest sub-profile containing at least one nodal parameter. Using Steps A1–A4, estimate all parameters in the sub-profile. Repeat Step B2 until all parameters in the profile are estimated.