NONLINEAR FILTERS IN GENOMIC CONTROL

Edward R. Dougherty,[1] Michael L. Bittner,[2] Yidong Chen,[2] Seungchan Kim,[1]
Krishnamoorthy Sivakumar,[3] Junior Barrera,[4] Paul Meltzer,[2] Jeffery M. Trent[2]

[1] Department of Electrical Engineering, Texas A&M University
[2] Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health
[3] Department of Electrical Engineering, Washington State University
[4] Departmento de Ciencia da Computacao, Universidade de São Paulo

ABSTRACT

The genome is a highly complex nonlinear control system regulating cell function. One of the primary means for regulating cellular activity is the control of protein production. Protein production is controlled by the amount of mRNA expressed by individual genes. This level of gene expression is modulated by protein machinery that senses conditions internal and external to the cell. A set of gene expression levels over time can be modeled as a random time function. The tools required to start building an understanding of genomic regulation of expression are those which allow one to discern the probability characteristics of this random function. Basic to such understanding is the ability to discover how expression levels of various genes predict other expression levels, both at fixed moments in time and through time. Hence, genomic pathway analysis depends on the application of nonlinear signal filters. In the approach here, gene expression levels can have three logical values: −1 (down-regulated), 1 (up-regulated), 0 (invariant). Filters are in terms of ternary logic and are statistically optimized via conditional probabilities. Although recent cDNA microarray technology permits simultaneous measurement of thousands of gene expression levels, current technology severely limits the number of microarrays one can use to acquire data for filter design. Consequently, one is required to constrain optimization by using small numbers of predictor genes or restricting the filter class. The present paper gives an overview of the general paradigm for genomic prediction and provides some early results.

1. INTRODUCTION

Estimates of the total number of genes in the human genome range from 70,000 to more than 100,000. As a result of the Human Genome Project's efforts to identify the complete complement of human genes, sequences and clones for more than 1.3 million expressed sequence tagged sites (ESTs) are currently publicly available. The sequences in the databases can be clustered into 62,851 non-redundant clusters (genes). This set of gene tags may represent half or more of the human gene set. It is obvious that in order for an organism to function, each gene must be expressed in a specific temporal and spatial context, thereby providing its unique function. The ability to link these genetic sequences to their functions has fallen far behind the ability to collect them, as only 14% of the identified clusters contain genes associated, even tenuously, with a known functionality.

Recently, methods for carrying out genome-wide expression studies have been described, in particular, cDNA microarrays [1-6]. For the latter, cDNA clone inserts are robotically printed onto a glass slide and subsequently hybridized to two differently fluorescently labeled cDNA representations of total RNA pools from test and reference cells, thereby allowing one to determine the relative abundance of transcript present in the pool. To gain insight into a gene's functional behavior in a cell, it is necessary to study a gene expression pattern (either in a temporal setting, in different cellular contexts, or under various circumstances) as it responds to the environment and to the action of other genes. Such experiments produce much larger data sets concerning transcriptional levels than have previously been available for analysis. To date, only relatively simple analysis, based on correlation of changes, has been applied to these data sets. As it has become clear that the control of transcription is accomplished by a method utilizing a variety of inputs [7, 8], it is necessary to design analytical tools that can sift expression profile data to detect the types of multivariate influences on decision-making operating within complex genetic networks. This paper discusses the general paradigm for gene expression prediction within the context of nonlinear filtering and provides some early results

To survey the association between genes, enough differing conditions must be sampled so that the independent functioning of different genetic networks is accessed. The amount of sampling requires data from numerous cDNA microarrays. To standardize expression measurements derived from different arrays, a consistent reference sample is simultaneously hybridized to each array with various test samples. The expression levels are reported as ratios relative to the reference sample. Further reliability can be obtained by using the variance of a large set of "housekeeping" genes to estimate the statistical significance of observed ratios. An algorithm calibrates the data internally to each microarray and statistically determines whether the data justifies the conclusion that an expression is up-regulated or

down-regulated with 99% confidence [9]. As a first step in carrying out nonlinear gene expression prediction on gene expression profiles, data complexity is reduced by thresholding the changes in transcript level into ternary expression data: [–1 (down-regulated), +1 (up-regulated), or 0 (invariant)].

From a statistical perspective, the mechanism of the association is not a factor, only the ability to predict the target level from the predictor levels. The predictor genes may be upstream or downstream from the target gene in the actual genetic network, some may be upstream and some downstream, or they may be distributed about the genetic network in such a way that their relation to the target gene is based on chains of interaction of various intermediate genes. Thus, whatever the relationship of the predicting genes to the predicted, if knowledge of their states allows us to better predict the expression level of the target gene, then we infer there is some relationship — the better the prediction, the stronger the relation. Given that a goal of human genomics research is to begin elucidating the function of newly discovered genes, the ability of the method to find connections independent of exact knowledge of all of the components of a gene network or even of the exact form of their interaction is a key strength of the method.

2. GENE EXPRESSION AS A RANDOM FUNCTION

The mRNA being measured by a microarray comes from a large collection of cells within a paticular cell line and therefore the expression level represents an expression average across the cell line at a moment in time. If we let $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ denote the vector of gene expression levels on a microarray, as a function of time t, $\mathbf{X}(t)$ is a vector random function. Genetic changes over time concern the random process $\mathbf{X}(t)$ as a random function of t; questions regarding the interrelation between genes at a given moment of time concern $\mathbf{X}(t_0)$ for a fixed t_0. Comparison of two cell lines, say tumerogenic and nontumerogenic, involves two random processes $\mathbf{X}_1(t)$ and $\mathbf{X}_2(t)$, and their cross probabilistic characteristics. For instance, if one is only concerned with their second-order characteristics, then their cross-covariance function is of interest.

The genome is not a closed system. It is affected by intracellular activity, which in turn is affected by external factors. At a very general level, we might represent the situation by a pair $(\mathbf{X}(t), \mathbf{Z}(t))$, where $\mathbf{Z}(t)$ is a vector random function of time-dependent random variables external to the genome, either cellular or otherwise. In any practical situation, $\mathbf{Z}(t)$ will only include variables that can be observed, are measurable, and are of interest. In a laboratory setting, $\mathbf{Z}(t)$ might be a deterministic function composed of one or two components decided upon by the experimenter. For instance, if we are interested in cellular reaction to heat shock at time t_0, then we might let $Z(t)$ be the deterministic function having unit-impulse at t_0 and being null otherwise. The effect of the heat shock after a time increment Δt is described by the transition $\mathbf{X}(t_0) \rightarrow \mathbf{X}(t_1)$, where $t_1 = t_0 + \Delta t$. Had we a complete stochastic model of the biological mechanisms set into motion by heat shock and were we able to nullify other external effects, then we could fully describe this transition by the joint probability distribution for $\mathbf{X}(t_0)$ and $\mathbf{X}(t_1)$

$$F_{\mathbf{X}}(\mathbf{x}_0, \mathbf{x}_1; t_0, t_1) = P(\mathbf{X}(t_0) \le \mathbf{x}_0, \mathbf{X}(t_1) \le \mathbf{x}_1)$$

More generally, we might only know that one of several randomly selected external conditions has been imposed, in which case the transition $\mathbf{X}(t_0) \rightarrow \mathbf{X}(t_1)$ is still our concern; however, since $\mathbf{Z}(t)$ is now a random function, $F_{\mathbf{X}}(\mathbf{x}_0, \mathbf{x}_1; t_0, t_1)$ is a marginal distribution relative to the joint probability distribution $F_{\mathbf{X},\mathbf{Z}}((\mathbf{x}_0, \mathbf{z}_0), (\mathbf{x}_1, \mathbf{z}_1); t_0, t_1)$. At present, we are a long way from full characterization and are interested in partial characterization of $F_{\mathbf{X}}(\mathbf{x}_0, \mathbf{x}_1; t_0, t_1)$ and in using partial characterization to obtain insight into genomic signaling pathways.

As with any signal processing setting, the most critical problem is prediction of $\mathbf{X}_1 = \mathbf{X}(t_1)$ from observation of $\mathbf{X}_0 = \mathbf{X}(t_0)$. Relative to mean-square error (MSE), the best predictor is the conditional expectation $E[\mathbf{X}_1|\mathbf{X}_0]$. Given the quantized setting, we can proceed in the manner used for the design of morphological operators [10-12]. The degree to which a designed predictor approximates the optimal predictor depends on the training procedure and the sample size. Even for a relatively small number of predictor genes, good design requires a data set sufficiently large to obtain precise estimates of the 3^n conditional expectations $E[\mathbf{X}_1|\mathbf{x}_0]$. However, we typically have a small number of microarrays, less than 100. The error, ε_{app}, of a designed approximation of the optimal predictor is a sum of two errors, $\varepsilon_{app} = \varepsilon_{opt} + \varepsilon_{des}$, where ε_{opt} is the error of the optimal predictor and ε_{des} is the design error resulting from estimating the predictor from the data. As $N \rightarrow \infty$, $\varepsilon_{des}(N) \rightarrow 0$, but for the small numbers used in practice, ε_{des} is significant [13].

The data problem can be mitigated if, rather than estimating the best predictor from a too small data set, we estimate the best predictor from a constrained set of predictors. Its theoretical error will exceed that of the best unconstrained predictor; however, it can be designed more precisely from the data. The error, $\varepsilon_{app,C}$, of a designed approximation of the best predictor satisfying a constraint C is the sum of two errors, $\varepsilon_{app,C} = \varepsilon_{opt,C} + \varepsilon_{des,C}$, where $\varepsilon_{opt,C}$ is the error of the optimal predictor and $\varepsilon_{des,C}$ is the design error. The dilemma of finding good predictors of gene expression levels is threefold: (1) $\varepsilon_{opt} < \varepsilon_{opt,C}$ and $\varepsilon_{des} > \varepsilon_{des,C}$; (2) ε_{opt} is decreased by using more predictor genes but ε_{des} increases with more predictor genes; (3) the stronger the constraint, the more $\varepsilon_{des,C}$ is reduced, but at the cost of increasing $\varepsilon_{opt,C}$. If we had access to an unlimited number of microarrays, then we could make both ε_{des} and $\varepsilon_{des,C}$ arbitrarily small and have $\varepsilon_{app} \approx \varepsilon_{opt} < \varepsilon_{opt,C} \approx \varepsilon_{app,C}$. In our low-replication environment, ε_{des} can greatly exceed $\varepsilon_{des,C}$. Consequently, the error of the designed constrained predictor can be significantly smaller than that of the designed unconstrained filter. Choosing an appropriately strong constraint is one of the key problems of nonlinear filter theory [14].

To discover predictive relationships in the gene set, we let each gene be the target and design predictors for each combination of K or less predictor genes. If a designed predictor yields a low error, then we conclude that the predictor genes possess a pre-

dictive relation with the target. For design and testing, the total microarray sample is randomly split into M training sets and $N - M$ test sets. The predictor is designed on the M training data sets and then applied to the $N - M$ test sets to obtain a test error, which serves as an estimate of the population error for the designed predictor. To obtain good error estimates, this procedure is repeated 256 times and the estimated error, ε_{test}, is taken as the average of these errors. For the optimal nonlinear predictor, ε_{test} serves as an estimate of ε_{app}; for an optimal constrained predictor, ε_{test}, which we write as $\varepsilon_{test,C}$, estimates $\varepsilon_{app,C}$. The errors for the optimal unconstrained and constrained predictors are $\varepsilon_{opt} = \varepsilon_{test} - \varepsilon_{des}$ and $\varepsilon_{opt,C} = \varepsilon_{test,C} - \varepsilon_{des,C}$, respectively. Since we do not know ε_{des} and $\varepsilon_{des,C}$, we take ε_{test} and $\varepsilon_{test,C}$ as measures of goodness for the designed unconstrained and constrained nonlinear predictors, respectively. We must keep in mind that ε_{des} always exceeds, and can greatly exceed, $\varepsilon_{des,C}$. Hence, if ε_{test} is close to $\varepsilon_{test,C}$, then it is reasonable to expect that the optimal unconstrained nonlinear predictor significantly outperforms the optimal constrained predictor, meaning that ε_{opt} is significantly less than $\varepsilon_{opt,C}$. Nonetheless, the only actual empirical measures of performance we have are ε_{test} and $\varepsilon_{test,C}$, and these are based on the 256 twenty-training/ten-test data splits.

We use the class of ternary perceptrons as our set of constrained predictors. For predicting a target Y from a collection of predictor expressions X_1, X_2, ..., X_n, a ternary peceptron is of the form

$$Y_{pred} = T(a_1X_1 + a_2X_2 + \cdots + a_nX_n + b)$$

where $T(z) = -1$ if $z < -0.5$, $T(z) = 0$ if $-0.5 \le z \le 0.5$, and $T(z) = +1$ if $z > 0.5$. Perceptron design requires estimating the coefficients a_1, a_2,..., a_n, and b from microarray data, and we have done this using a stochastic training algorithm [15].

Whether using constrained or unconstrained predictors, for a target expression Y, we take the MSE of Y_{pred} and normalize this MSE by dividing by the MSE of $T(\mu_Y)$, which is the error resulting from taking the threshold of the mean of Y as the predictor of Y. This yields the normalized mean-square error $NMSE[Y_{pred}]$. We measure the degree of prediction by $NMSE[Y_{pred}]$ or, equivalently, by

$$R^2 = \frac{MSE[T(\mu_Y)] - MSE[Y_{pred}]}{MSE[T(\mu_Y)]} = 1 - NMSE[Y_{pred}]$$

Since $NMSE[Y_{pred}] \le 1$, $0 \le R^2 \le 1$. The closer $NMSE[Y_{pred}]$ is to 0, the better Y is predicted, and the more Y is determined by X_1, X_2,..., X_n via the predictor. We use the NMSE for both perceptron and unconstrained nonlinear predictors to measure the strength of the predictive relation between a set of predictor genes and a target gene.

3. Experimental Results

In our first study using these methods, we have examined predictive relationships for a certain set of genes in the context of their responsiveness to genotoxic stresses [15]. We used data from a microarray study surveying transcription of 1238 ESTs during the response of a myeloid line to ionizing radiation [16]. In the study, 30 genes not previously known to participate in response to IR were found to be responsive. To further characterize the responsiveness of these genes to genotoxic stresses, the responsiveness of a subset of nine of them was examined by blot assays in 12 cell lines stimulating with ionizing radiation, a chemical mutagen (methyl methane sulfonate, MMS), or ultraviolet radiation. The ternary expression levels for 12 genes and their 3 external experimental conditions are given in Table 1. For experimental methods and the full discussion related to Table 1, see Ref. [16]. In this paper we will not go into great detail on the results of our analysis, but will focus instead on a few results of interest relative to nonlinear signal processing.

Table 1.

	Genes												Condition		
Cell-line	RCH1	BCL3	FRA1	REL-B	ATF3	IAP-1	PC-1	MBP-1	SSAT	MDM2	P21	P53	IR	MMS	UV
ML-1	-1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
ML-1	0	0	0	0	1	0	0	0	0	1	1	1	0	1	0
Molt4	-1	0	0	1	1	0	1	0	0	1	1	1	1	0	0
Molt4	0	0	1	0	1	0	0	0	0	0	1	1	0	1	0
SR	-1	0	0	1	1	1	1	1	0	1	1	1	1	0	0
SR	0	0	0	0	1	0	0	0	0	1	1	1	0	1	0
A549	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0
A549	0	0	0	0	1	0	0	0	0	0	1	1	0	1	0
A549	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1
MCF7	-1	0	1	1	0	0	0	0	0	1	1	1	1	0	0
MCF7	0	0	1	0	1	0	0	0	0	1	1	1	0	1	0
MCF7	0	0	1	1	1	0	0	0	0	1	1	1	0	0	1
RKO	0	1	0	1	1	1	1	0	0	1	1	1	1	0	0
RKO	0	0	0	0	1	0	0	0	0	0	1	1	0	1	0
RKO	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1
CCRF-CEM	-1	1	1	1	1	0	1	0	0	0	0	-1	1	0	0
CCRF-CEM	0	0	0	0	1	0	0	0	0	0	0	-1	0	1	0
HL60	-1	1	0	1	1	0	1	0	1	0	1	-1	1	0	0
HL60	0	0	1	0	1	0	0	0	0	1	1	-1	0	1	0
K562	0	0	0	0	0	0	0	0	0	0	0	-1	1	0	0
K562	0	0	0	0	1	0	0	0	0	0	0	-1	0	1	0
H1299	0	0	0	1	0	0	1	0	0	0	0	-1	1	0	0
H1299	0	0	0	0	1	0	0	0	0	0	1	-1	0	1	0
H1299	0	0	0	0	1	0	1	0	0	0	1	-1	0	0	1
RKO/E6	-1	1	0	1	0	1	1	0	0	0	0	-1	1	0	0
RKO/E6	-1	0	0	0	1	0	0	0	0	0	1	-1	0	1	0
RKO/E6	-1	0	0	0	1	0	0	0	0	0	1	-1	0	0	1
T47D	0	0	0	1	0	0	0	0	0	0	1	-1	1	0	0
T47D	0	0	0	0	1	0	0	0	0	0	1	-1	0	1	0
T47D	0	0	0	0	1	0	0	0	0	0	1	-1	0	0	1

From a filter-design perspective, the most interesting aspect of the problem is the large number of random variables and the extreme paucity of data. Our methodology is designed to discover nonlinear predictive relations, rather than design specific optimal filters. Each time the data is split between training and testing, a best filter is determined and, as the procedure is repeated 256 times, many different best filters arise. For a given

gene-expression predictor set and target, ε_{test} gives the average error and ε_{test} provides a coefficient of determination, the provision of which is our main intent. Should we wish to estimate the optimal filter from the full data set, then we would use all of the microarrays for training, but this would leave no way to estimate filter error.

To the extent that a perceptron provides an approximation to a linear filter, if the optimal predictor is a perceptron, one might conclude that there is a somewhat linear relation between the predictor variables and the target. Because $\varepsilon_{des\text{-}per}$ is often substantially less than ε_{des}, it is often the case that $\varepsilon_{test\text{-}per} < \varepsilon_{test}$ even though $\varepsilon_{opt\text{-}per} > \varepsilon_{opt}$. For the genomic-stress-response data, for predictor genes IAP-1, PC-1, and SSAT, and target gene BCL3, the test errors are $\varepsilon_{test\text{-}per} = 0.336$ and $\varepsilon_{test} = 0.666$, which means that $\varepsilon_{des} - \varepsilon_{des\text{-}per} \geq 0.330$. Nevertheless, the inherent nonlinearity of genomic regulation can be sufficiently strong to overcome the estimation-error differential. For target BCL3, $\varepsilon_{test\text{-}per} = 0.826$ and $\varepsilon_{test} = 0.493$ for predictors RCH1, PC-1, and p53. In another striking example, gene REL-B is predicted with $\varepsilon_{test\text{-}per} = 0.472$ and $\varepsilon_{test} = 0.397$ using the predictor genes BCL3 and ATF3, in conjunction with external application of ionizing radiation.

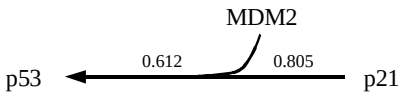

Figure 1.

For presentation, prediction results are shown as arrow plots, with the target gene at the left, and the chained predictor plotted to the right. The NMSE achieved by the adjoining a predictor gene is placed on the arrow preceding it. Many predictions produced by perceptrons confirm biological expectation. For example, it is known that p53 is influential, but not determinative of the up regulation of both p21 and MDM2. Thus, some level of prediction is possible by a combination of these two genes. This expectation is clearly met, as shown in Fig. 1.

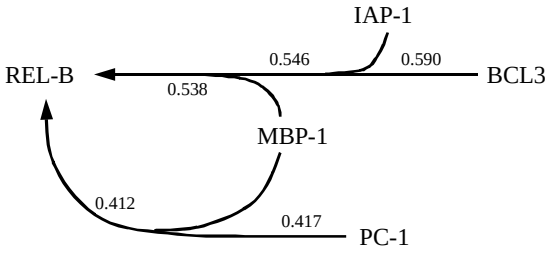

Figure 2.

For a set of newly found IR responsive genes [REL-B, RCH1, PC1, MBP-1, BCL3, IAP-1], perceptron prediction indicates their expression behaviors are interrelated. For example, REL-B is known to be a modifier of NFκB, a transcription activator commonly induced in response to genotoxic shock. As shown in the prediction tree in Fig. 2, it is linked to MBP-1 (a zinc finger protein that binds to a variety of enhancer elements), PC1 (a subtilisin-like proprotein processing enzyme found to be frequently highly expressed in carcinoid tumors), BCL3 (a B-cell lymphoma 3 encoded protein that is a putative transcription activating factor whose expression is mitogenically stimulated), and IAP-1 (an inhibitor of cellular apoptosis). It is known that these genes do not constitute a simple functional pathway. What can be seen to be common in the transcription pattern of these proteins is the stimulus, ionizing radiation, to which they respond directly or indirectly. This effect is clear if the experimental conditions (IR, MMS and UV) are included into the set of predictors. The results for REL-B are given in Fig. 3, where the IR condition becomes the sole source of prediction. Additions of further genes do not increase the accuracy of prediction significantly.

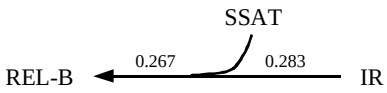

Figure 3.

4. CONCLUSION

Viewing the genome as a control system, receiving and giving signals both internally and externally, it is natural to model gene expression as a vector random function over time. Quantization according to whether a gene is up-regulated, down-regulated, or invariant leads at once to nonlinear digital signal processing as the appropriate framework to view gene expression prediction. Hence, we apply the methods of statistical filter optimization, and these can be applied in the context of computational mathematical morphology. This paper has presented the framework and demonstrated the statistical methodology for genes undergoing genotoxic stress.

5. REFERENCES

[1] De Risi, J. L., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and J. M. Trent, "Use of a cDNA microarray to analyse gene expression patterns in human cancer," *Nature Genetics*, **14**, 457-460, 1996.

[2] De Risi, J. L., Iyer, V. R., and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, **278**, 680-686, 1997.

[3] Schena, M., Shalon, D., Davis, R. W., and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, **270**, 467-470, 1995.

[4] Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., and R. W. Davis, "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes," *Proc. National Academic Science, USA*, **93**, 10614-10619, 1996.

[5] Duggan, D.J., Bittner, M. L., Chen, Y., Meltzer, P. S., and J. M. Trent, "Expression profiling using cDNA microarrays," *Nature Genetics*, **21**, 10-14, 1999.

[6] Brown, P. O., and D. Botstein, "Exploring the new world of the genome with DNA microarrays," *Nature Genetics*, **21**, 33-37, 1999.

[7] McAdams, H. H., and L. Shapiro, "Circuit simulation of genetic networks," *Science*, **269**, 650-656, 1995.

[8] Yuh, C. H., Bolouri, H., and E. H. Davidson, "Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene," *Science*, **279**, 1896-1902, 1998.

[9] Chen, Y., Dougherty, E. R., and M. L. Bittner, "Ratio-based decisions and the quantitative analysis of cDNA microarray images," *Biomedical Optics*, **2**, 364-374, 1997.

[10] Dougherty, E. R., and R. P. Loce, "Optimal mean-absolute-error hit-or-miss filters: morphological representation of the conditional expectation," *Optical Engineering*, **32**, 815-827, 1993.

[11] Dougherty, E. R., and D. Sinha, "Computational mathematical morphology," *Signal Processing*, **38**, 69-85, 1995.

[12] Barrera, J., Dougherty, E. R., and N. S. Tomita, "Automatic programming of binary morphological machines by design of statistically optimal operators in the context of computational learning theory," *Electronic Imaging*, **6**, 54-67, 1997.

[13] Dougherty, E. R., and R. P. Loce, "Precision of morphological-representation estimators for translation-invariant binary filters: increasing and nonincreasing," *Signal Processing*, **40**, 129-154, 1994.

[14] Dougherty, E. R., and J. Barrera, "Logical image operators," *Nonlinear Filters for Image Processing*, E. R. Dougherty and J. Astola, eds., SPIE and IEEE Presses, Bellingham, 1999.

[15] Kim, S., Dougherty, E. R., Bittner, M. L., Sivakumar, K., Chen, Y., Kamat, V., Meltzer, P., and J. M. Trent, "Nonlinear multivariate analysis of gene interaction via expression prediction," (submitted).

[16] Amundson, S., Bittner, M. L., Chen., Y., Trent, J. M., Meltzer, P., and A. J. Fornace, "cDNA microarray hybridization reveals complexity and heterogeneity of cellular genotoxic stress responses," (submitted).