Binarization of Microarray Data on the Basis of a Mixture Model[1]

Xiaobo Zhou, Xiaodong Wang, and Edward R. Dougherty[2]

Department of Electrical Engineering, Texas A&M University, College Station, Texas, 77843 [X. Z., E. R. D.]; Department of Electrical Engineering, Columbia University, New York, New York 10027 [X. W.]; and Department of Pathology, University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030 [E. R. D.]

Abstract

Although gathered as continuous data, expression measurements from gene microarrays may be quantized before downstream analysis and modeling. This is especially true for modeling gene prediction and genetic regulatory networks. Coarse quantization results in lower computational requirements, lower data requirements for model inference, and easier conceptualization. This paper proposes a mixture model for binarization. For each gene, the model, composed of a sum of two distributions, is fit to expression data for that gene, and data points are binarized according to the model. The mixture model is based on the assumption of multiplicative up-regulation. The proposed method is compared with mean and median binarization by comparing classification performance based on the binary data from the different methods. Classification is performed for simulated data generated from a microarray model studied previously and for cancer data arising from two studies involving hereditary breast cancer and small, round blue-cell tumors of childhood.

Introduction

It is a general principle that a system should be modeled at the lowest level of complexity that permits the accomplishment of the purposes for which the model is being developed. Lower levels of model complexity mean less computation, lower data requirements for model identification, and greater ease of conceptualization. In a sense, this is an engineering form of Occam's razor in which the pragmatics of the problem imply some minimal level of necessary complexity. The issue of complexity reduction is particularly salient for the development of prediction models and genetic regulatory models using microarray data, because the number of genes is very large and the number of samples very small. Modeling involves numerous goals, including prediction of targets based on pathways, characterization of disease states, and the design of optimal time-dependent dosing regimens (1). For gene-expression-based models, various modeling decisions must be made based on ones goals, computational power, and data abundance: the set of genes to be included in the model, the degree of complexity allowed for functional relations between genes within the model, and the quantization of expression levels.

Regarding quantization, the subject of this paper, perhaps the most well-studied genetic regulatory network model is the Boolean network (2–4). As the name implies, this model uses binary quantization in which a gene is either ON (expression level equals 1) or OFF (expression level equals 0). The intent of such a coarse-grained model is not to serve as an architecture for biochemical pathways, but rather to give qualitative insight into multivariate, dynamical gene interaction, to provide a mathematical structure to study this dynamical behavior at the level of logical switching, and to design therapeutic strategies based on the dynamical behavior of the network when the ON-OFF paradigm is sufficient. Application of Occam's razor is implicit in the study of dynamical behavior and the design of therapies in the context of binary quantization; indeed, both are valid to the extent that they can be pragmatically described in a binary framework. In fact, the Boolean model has provided conceptual insights into the behavior of genetic regulatory networks (5–7). The Boolean-network model, which is deterministic, has been extended recently to a stochastic framework that allows for different functional relationships and perturbations between states. These stochastic networks are called "probabilistic Boolean networks" (8). A probabilistic Boolean network maintains the rule-based structure of a Boolean network, whereas allowing for uncertainty. Two studies have demonstrated that intervention can be addressed within the context of probabilistic Boolean networks (9, 10), and a third has considered optimal time-dependent external control based on dynamic programming (11). Even should the necessary technologies for diagnosis, monitoring, and therapy for these kinds of model-based strategies become practically feasible, their successful use would still depend on appropriate binarization of continuous data. The issue of binarization has been addressed in some detail in a recent paper, and we refer to it for a more in-depth discussion of binarization issues (12).

The approach we take in this paper is to apply a binarization procedure based on a multiplicative model for expression up-regulation. The expression level of a gene varies across a set of microarrays, and in the context of a binary quantization is either ON or OFF for the various samples yielding the microarrays. Taken as a collection, the meas-

Received 2/5/03; revised 4/14/03; accepted 4/16/03.
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.
[1] Supported by the National Human Genome Research Institute, the University of Texas M. D. Anderson Cancer Center, and the National Science Foundation grant DMS-0225692.
[2] To whom requests for reprints should be addressed, at Department of Electrical Engineering, Texas A&M University, College Station, TX 77843. E-mail: edward@ee.tamu.edu.

urements for a particular gene compose a distribution of values, and we shall model that distribution as a mixture of two distributions, one corresponding to lack of up-regulation and the other to up-regulation. The parameters of these two distributions will be estimated from the expression data for the gene, and then data will be quantized according to the modeled distributions. The mixture model depends on whether one is using ratios or the intensities directly. We focus on ratios, and comment on how the procedure applies to straight intensity data with a simpler model.

Materials and Methods

Mixture Model. The mixture model used for binarization is based on the assumption of multiplicative up-regulation. For a particular gene, g, we assume that its values across the set of microarrays for which it is not up-regulated are described by a normal random variable U of which the mean is the nominal value of g in these samples. The model has a multiplicative factor $K > 1$ such that the values of g in the up-regulated samples are governed by the random variable KU, which is also normal because U is normal. Assuming a two-channel cDNA microarray, the values of the reference channel are modeled by a random variable B. Taking logs, there are two possibilities. For samples in which g is not up-regulated, we get:

$$\log \frac{U}{B} = \log U - \log B. \tag{A}$$

For an up-regulated sample, we get:

$$\log \frac{KU}{B} = \log K + \log U - \log B. \tag{B}$$

If we make the simplifying assumption that B is not random, then the distribution of log KU/B is simply a shift of the distribution of log U/B and both possess a distribution that is the log of a normal distribution. Hence, across the samples, the logs of the ratios will appear as a mixture of two log-of-normal (not lognormal) distributions.

On the basis of the preceding considerations, we postulate a log-of-normal mixture model for the observations of a particular gene across a set of microarrays. We assume that the logs follow the mixture model

$$X \sim \sum_{k=1}^{Q} c_k \mathcal{LN}(\mu_k, \sigma_k^2), \tag{C}$$

where X is the log of the ratio, $\mathcal{LN}$ denotes the log-of-a-normal distribution having density

$$\phi(X; \mu_k, \sigma_k^2) \triangleq \frac{2e^X}{\sqrt{2\pi}\sigma_k} \exp\left\{ - \frac{[e^X - \mu_k]^2}{2\sigma_k^2} \right\}, \tag{D}$$

and c_k, μ_k and σ_k^2 are the weights, means, and variances, respectively, of the mixture model. Q is the number of quantization levels, which in our case is $Q = 2$, but could be different were we to model more levels of multiplicative up-regulation. Estimation of the model parameters is explained in "Appendix." Fig. 1 shows a mixture of two log-of-normal distributions.

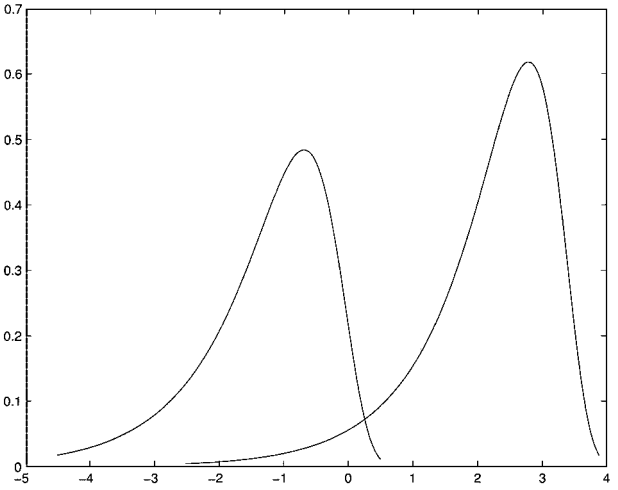

Fig. 1. A mixture of two log-of-normal distributions.

Binarization based on the log-of-normal mixture model is achieved by thresholding. Without loss of generality, assume $\mu_1 < \mu_2$. The log of a ratio X is quantized by:

$$\begin{aligned} q(X) &= \begin{cases} 0, & X \le T, \\ 1, & X > T, \end{cases} \\ \text{with } T &\triangleq \frac{\bar{\mu}_1 + \bar{\sigma}_1 + \bar{\mu}_2 - \bar{\sigma}_2}{2}, \end{aligned} \tag{E}$$

where for $i = 1, 2$,

$$\begin{aligned} \bar{\mu}_i &\triangleq \int_{-\infty}^{\infty} X \phi(X; \mu_i, \sigma_i^2) dX \\ &= \int_0^{\infty} \frac{2}{\sqrt{2\pi}\sigma_i} \exp\left(- \frac{z^2}{2} \right) \ln(\mu_i + \sigma_i z) dz \\ \text{and} \quad \sigma_i^2 &\triangleq \int_{-\infty}^{\infty} (X - \bar{\mu}_i)^2 \phi(X; \mu_i, \sigma_i^2) dX. \end{aligned} \tag{F}$$

Because closed forms do not exist for the preceding integrals, they are evaluated by Monte Carlo methods.

Were we to use direct intensities instead of ratios, then we would not have quotients and the not-up-regulated and regulated cases would simply involve a normal random variable U and a second normal random variable KU, respectively. Hence, we would apply a Gaussian mixture model.

Testing Binarization Performance Using a Simulation Model. We test the performance of MMB[3] using a model-based simulation and compare it to using the mean and median of the samples for each gene, denoted by "Mean" (12) and "Median" (12, 13), respectively. The Mean method is based on a threshold T (*i.e.,* replacing the T in E) that is the mean of the intensities of each gene, and the Median method

[3] The abbreviations used are: MMB, mixture model binarization; SRBCT, small, round blue-cell tumor; NB, neuroblastoma; RMS, rhabdomyosarcoma.

is based on a threshold T that is the median of the intensities of each gene. Then the intensity of this gene is quantized to 1 if the intensity exceeds T, and it is 0 otherwise. The latter two methods are simple but effective if there is sufficient separation in the mixture.

The simulated data are generated by a parameterized random signal model (14). It is assumed that the n genes on the m microarrays have the mean expression levels $I_1, I_2, \ldots, I_n$, with $I_i = 100 + \xi_i$, where ξ_i follows an exponential distribution with mean 3000. The model assumes that the intensities U_{i1}, $U_{i2}, \ldots, U_{im}$ for a specific gene g_i follow a normal distribution $\mathcal{N}(I_i, (\alpha I_i)^2)$, where α is a fixed coefficient of variation. Here we set $\alpha = 0.2$, which is realistic. With the inclusion of normally distributed additive noise N_{ij}, observed intensities are given by:

$$Z_{ij} = U_{ij} + N_{ij}, \quad i, = 1, \ldots, n, \quad j = 1, \ldots, m, \qquad \text{(G)}$$

for samples in which gene g_i is not up-regulated, and

$$Z_{ij} = KU_{ij} + N_{ij}, \quad i, = 1, \ldots, n, \quad j = 1, \ldots, m, \qquad \text{(H)}$$

when gene g_i is up-regulated. The model assumes that the additive noise varies from microarray to microarray, with $N_{ij} \sim \mathcal{N}(0, \eta_i^2)$, where $\eta_i \sim \mathcal{N}(\mu_a, \sigma_a^2)$. Here we set $\mu_a = 30$ and $\sigma_a = 10$. To complete the ratio model, we assume the reference channel has normally distributed values B_{ij} with mean equal to the mean, μ_I, of the intensity means and SD $\alpha\mu_I$. In essence, this means that we are assuming a uniform reference probe across all of the genes, and the mean probe intensity is normalized to the mean intensity of the expression means. Other reference models could be used, for instance, assuming each gene to have its own probe.

We test binarization performance by comparing the effects of the different methods on classification. On the basis of the simulation model, we take the log

$$X_{ij} = \log \frac{Z_{ij}}{B_{ij}}, \quad i, = 1, \ldots, n, j = 1, \ldots, m \qquad \text{(I)}$$

quantize the log to obtain $q(X_{ij})$, and then attempt to perform classification of the up-regulated and not-up-regulated samples based on gene expression. The testing principle is that binarization yielding good classification is a satisfactory binarization (at least for classification purposes). For classification we use a recently proposed method involving Bayesian variable selection in conjunction with a probit regression model to relate the gene expression with the class (15). This method has performed well on labeled data from hereditary breast cancer, and we will be able to compare those earlier results with those obtained from binary expression data.

Considering the high computational cost of the Bayesian gene selection, we use the principle that the smaller the sum of squares within groups and the larger the sum of squares between groups, the better the classification performance. Hence, we take the ratio of the between-group to within-group sum of squares and determine a threshold so that we only keep those genes for which the ratio exceeds the threshold. The leave-one-out method is used to estimate classification errors.

Table 1 Recognition accuracy (%)

K	MMB	Median	Mean
1.5	83.74	78.41	78.17
1.8	87.53	84.54	83.65
2.0	89.53	87.54	85.94
2.5	96.89	96.82	95.05

Subsequent to the simulation, we consider tumor classification using two different published data sets. Once again we are interested in which binarization method yields the best classification rate. It will be seen that MMB has no errors for both data sets. Not only does this show the worth of using the mixture model (which is our main intent), but it also lends support to a proposition put forth by Shmulevich and Zhang (12) that binary data can often perform important data analysis tasks. There, the authors demonstrate the ability of binary data to form meaningful clusters; here, the experiments demonstrate the ability of binary data to perform successful classification.

The cancer data will be normalized to account for variation between microarrays. For the jth array we take the log transform of the ratios $t_{1j}, t_{2j}, \ldots, s, t_{nj}$ to obtain log t_{ij}, for $i = 1, 2, \ldots, n$. The mean, m_j, of the log-transformed ratios for the jth array is computed, and we define the normalized ratio according to the equation

$$\log \bar{t}_{ij} \triangleq \log t_{ij} - m_j \qquad \text{(J)}$$

for $i = 1, 2, \ldots, n$. The normalized ratios are used for classification.

The sensitivity of the Bayesian-variable-selection method was examined in the original study by adding synthetic Gaussian noise to the data (15). Here we will check sensitivity using only actual data by considering how the binarization methods perform when the data are not normalized for microarray-induced variation. This has the effect of making the classification more difficult because it tends to increase the dispersion of expression measurements for each gene.

Results and Discussions

Simulation Study. The Bayesian selection algorithm ranks a gene according to the percentage of times it appears among the posterior samples generated by the algorithm (15). The five strongest (highest ranked) genes are used for classification for each binarization method being tested. These will differ according to the method. In each case, the five strongest genes are used in the probit classification to classify the up-regulated samples. The average recognition accuracies, based on 50 simulations, are shown in Table 1 for different up-regulation factors. There is little difference for large K, but when K is small, the mixture-model method significantly outperforms the Mean and Median methods. Because classification accuracy depends on K in the up-regulation model, this means that the mixture-model approach should help for more difficult classification problems.

Hereditary Breast Cancer Data. First we consider hereditary breast cancer data, which can be downloaded from the web page of the original paper (16). In that paper, cDNA

Table 2 Recognition accuracy (no. of errors) for two cancer data sets

	MMB	Mean	Median
Breast cancer data with normalization	0	1	1
Breast cancer data without normalization	0	1	3
SRBCT with normalization	0	0	1
SRBCT without normalization	0	4	3

Table 3 Strongest genes selected from the quantized breast cancer data using the MMB method

No.	CloneID	Gene description
1	26184	Phosphofructokinase, platelet
2	44180	α-2-macroglobulin
3	309583	ESTs
4	30502	Reticulon 1
5	812227	Solute carrier family 9 (sodium/hydrogen exchanger), isoform 1
6	376516	Cell division cycle 4-like
7	137638	ESTs
8	823940	Transducer of ERBB2, 1 (TOB1)
9	204897	Phospholipase C, γ 2 (phosphatidylinositol-specific)
10	839736	Crystallin, α B

microarrays are used in conjunction with classification algorithms to show the feasibility of using differences in global gene expression profiles to separate BRCA1 and BRCA2 mutation-positive breast cancers. Twenty-two tumor samples from 21 patients are examined: 7 BRCA1, 8 BRCA2, and 7 sporadic. There are 3226 genes for each tumor sample. After binarization, we use the probit regression classifier with Bayesian gene selection to classify BRCA1 *versus* BRCA2 and sporadic.

Table 2 gives the leave-one-out error counts for the three binarization methods being tested and the four experiments under consideration, the two cancers with normalized and non-normalized data. The 5 top-ranked genes under Bayesian variable selection are used for classification for each binarization method.

Tables 3 through 5 give the top 10 scoring genes found by Bayesian variable selection for MMB, Median, and Mean, respectively, when the data has been normalized. Using the top 5 genes for classification for each binarization method, MMB yields no errors, whereas both Mean and Median result in one error each. Given that there are 22 microarrays, this represents a 5% improvement using MMB. This is significant, especially because MMB yields no errors at all.

By comparing the lists in Tables 3 through 5 with the list of the top 10 scoring genes in the original paper (15), shown in Table 6, we see that even with the compression of binarization, Bayesian selection may still find a number of similar significant genes. The top 10 scoring genes using MMB with normalized data includes 3 of the genes on the original list, all 3 being in the original top 5. Included among the 3 are TOB1 and phosphofructonkinase. Neither Mean nor Median has any of the original top 10 among their respective top 10.

For the non-normalized data, we see from Table 2 that the performance of both MMB and Mean remain the same, with MMB still perfect, but Median results in three errors, which is very poor given that MMB has no errors.

Table 4 Strongest genes selected from the quantized breast cancer data using the Median method

No.	CloneID	Gene description
1	812227	Solute carrier family 9 (sodium/hydrogen exchanger), isoform 1
2	35865	Annexin A6
3	204299	Replication protein A3 (14kD)
4	809981	Glutathione peroxidase 4 (phospholipid hydroperoxidase)
5	245198	KIAA0130 gene product
6	126412	Androgen receptor associated protein 54
7	48406	Hydroxysteroid (17-β) dehydrogenase 4
8	712848	Mitogen-activated protein-kinase activating death domain
9	814595	Protein kinase C binding protein 1
10	825577	Steroidogenic acute regulatory protein related

Table 5 Strongest genes selected from the quantized breast cancer data using the Mean method

No.	CloneID	Gene description
1	768370	Tissue inhibitor of metalloproteinase 3
2	290871	Integrin, α 3 (antigen CD49C, α 3 subunit of VLA-3 receptor)
3	83210	Complement component 8, β polypeptide
4	812227	Solute carrier family 9, isoform 1
5	204299	Replication protein A3 (14kD)
6	814595	Protein kinase C binding protein 1
7	825577	Steroidogenic acute regulatory protein related
8	126650	ESTs
9	139354	ESTs
10	809981	Glutathione peroxidase 4 (phospholipid hydroperoxidase)

Table 6 Strongest genes listed in [15]

No.	CloneID	Gene description
1	897781	Keratin 8
2	823940	Transducer of ERBB2, 1 (TOB1)
3	26184	Phosphofructokinase, platelet
4	840702	Selenophosphate synthetase; Human selenium donor protein
5	376516	Cell division cycle 4-like
6	47542	Small nuclear ribonucleoprotein D1 polypeptide (16kD)
7	366647	Butyrate response factor 1 (epidermal growth factor-response factor 1)
8	293104	Phytanoyl-CoA hydroxylase (Refsum disease)
9	28012	*O*-linked *N*-acetylglucosamine (GlcNAc) transferase
10	212198	Tumor protein p53-binding protein, 2

Small Round Blue-Cell Tumor Data. Now we apply binarization and probit regression classification with Bayesian feature selection to a published data set for SRBCTs of childhood, which includes NB, RMS, non-Hodgkin's lymphoma, and the Ewing family of tumors (17). We classify the rhabdomysosarcoma and NB tumors. The data set for the two cancers is composed of 2308 genes and 35 samples, 23 samples for RMS, and 12 samples for NB.

Tables 7 through 9 give the 10 strongest genes using Bayesian selection with MMB, Median, and Mean, respectively. For the top 5 genes on each list, the classification, and

Table 7 Strongest genes selected from the quantized blue-cell-tumor data using the MMB method

No.	CloneID	Gene description
1	325182	Cadherin 2, *N*-cadherin (neuronal)
2	36950	Phosphofructokinase, liver
3	1435862	Antigen identified by monoclonal antibodies 12E7, F21, and O13
4	786084	Chromobox homolog 1 (*Drosophila* HP1 β)
5	1434905	Homeo box B7
6	137535	Transcriptional intermediary factor 1
7	774502	Protein tyrosine phosphatase, nonreceptor type 12
8	486175	Solute carrier family 16 (monocarboxylic acid transporters), member 1
9	866702	Protein tyrosine phosphatase, nonreceptor type 13
10	166236	Glucose-6-phosphate dehydrogenase

Table 8 Strongest genes selected from the quantized blue-cell-tumor data using the Median method

No.	CloneID	Gene description
1	36950	Phosphofructokinase, liver
2	39796	3-Hydroxymethyl-3-methylglutaryl-Coenzyme A lyase
3	726236	Paired mesoderm homeo box 1
4	1434905	Homeo box B7
5	814260	Follicular lymphoma variant translocation 1
6	38471	Human cyclin G1 interacting protein (1500GX1) mRNA, complete cds
7	729964	Sphingomyelin phosphodiesterase 1, acid lysosomal
8	782193	Thioredoxin
9	207358	Solute carrier family 2 (facilitated glucose transporter), member 1
10	866702	Protein tyrosine phosphatase, nonreceptor type 13

Table 9 Strongest genes selected from the quantized blue-cell-tumor data using the Mean method

No.	CloneID	Gene description
1	325182	Cadherin 2, N-cadherin (neuronal)
2	36950	Phosphofructokinase, liver
3	39796	3-Hydroxymethyl-3-methylglutaryl-Coenzyme A lyase
4	823598	Proteasome (prosome, macropain) 26S subunit, non-ATPase, 12
5	770394	Fc fragment of IgG, receptor, transporter, alpha
6	1435862	Antigen identified by monoclonal antibodies 12E7, F21, and O13
7	295985	ESTs
8	377461	Caveolin 1, caveolae protein, 22kD
9	75254	Cysteine and glycine-rich protein 2 (LIM domain only, smooth muscle)
10	50359	Mannose phosphate isomerase

the recognition accuracies are shown in Table 2. Again, MMB yields no errors. Cadherin 2, which tops the significant gene list for MMB, is identified as an important gene in the original SRBCT study (17). In fact, it can perfectly classify RMS and non-Hodgkin's lymphoma by itself using the binary data. Even with perfect single-gene classification possible, Median still produces one error, although Mean does yield perfect classification, with cadherin 2 at the top of its list.

The situation is very different for the non-normalized data. Again, MMB identifies cadherin 2 as the strongest gene and produces no errors. However, both Mean and Median perform poorly, with four and three errors, respectively.

Appendix

This appendix describes parameter estimation for the log-of-normal mixture model. First, the samples are partitioned into Q clusters using fuzzy C-means clustering, and then initial parameters are estimated using the standard vector quantization method. We then estimate the parameters in C using the expectation-maximization algorithm by iterating the following steps:

$$\gamma_k(X) = \frac{c_k \phi(X;\, \mu_k,\, \sigma_k^2)}{\sum_{k=1}^{K} c_k \phi(X;\, \mu_k,\, \sigma_k^2)}, \text{ for all } X \in \Gamma_i, \tag{A1}$$

$$\beta_k = \sum_{X \in \Gamma_i} \gamma_k(X), \tag{A2}$$

$$c_k = \beta_k \Big/ \sum_{j=1}^{K} \beta_j, \tag{A3}$$

$$\mu_k = \sum_{X \in \Gamma_i} [\gamma_k(X) e^X]/\beta_k, \tag{A4}$$

$$\sigma_k^2 = \sum_{X \in \Gamma_i} \gamma_k(X)[e^X - \mu_k]^2/\beta_k, \tag{A5}$$

Γ_i denotes the observations of gene i.

We note that, if Q were not fixed at $Q = 2$ for binary quantization, we could find an optimal number of distributions in the model via the minimum description length principle by building a binary tree with Q leaves, each representing a distribution in the mixture model (18).

References

1. Somogyi, R., and Greller, L. D. The dynamics of molecular networks: applications to therapeutic discover. Drug Discovery Today, *6:* 1267–1277, 2001.

2. Kauffman, S. A. Metabolic stability and epigenesist in randomly constructed genetic networks. J. Theor. Biol., *22:* 437–467, 1969.

3. Glass, K., and Kauffman, S. A. The logical analysis of continuous nonlinear biochemical control networks. J. Theor. Biol., *39:* 103–129, 1973.

4. Huang, S., Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. Mol. Med., *77:* 469–480, 1999.

5. Somogyi, R., and Sniegoski, C., Modeling the complexity of gene networks: understanding multigenic and pleiotropic regulation. Complexity, *1:* 45–63, 1996.

6. Wuensche, A. Genomic regulation modeled as a network with basis of attractions. Pacific Symp. Biocomput., *3:* 89–102, 1998.

7. Thomas, R., Thieffry, D., and Kaufman, M. Dynamical behavior of biological regulatory networks - 1. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. Bull. Math. Biol., 57: 257–276, 1995.

8. Shmulevich, I., Dougherty, E. R., Kim, S., and Zhang. W. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics, *18:* 261–274, 2002.

9. Shmulevich, I., Dougherty, E. R., and Zhang, W. Gene perturbation and intervention in probabilistic Boolean networks. Bioinformatics, *18:* 1319–1331, 2002.

10. Shmulevich, I., Dougherty, E. R., and Zhang, W. Control of stationary behavior in probabilistic Boolean networks by means of structural intervention. Biol. Systems, *10:* 431–446, 2002.

11. Datta, A., Choudhary, A., Bittner, M. L., and Dougherty E. R. External control in Markovian genetic regulatory networks. Machine Learning, *52:* 169–181, 2003.

12. Shmulevich, I., and Zhang, W. Binary analysis and optimization-based normalization of gene expression data. Bioinformatics, *18:* 555–565, 2002.

13. Dudoit S., Yang, Y. H., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P., Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res., 30: e15(1–10), 2002.

14. Balagurunathan, Y., Dougherty, E. R., Chen, Y., Bittner, M. L., and Trent, J. M. Simulation of cDNA microarrays via a parameterized random signal model. Biomed. Optics, *7:* 507–523, 2002.

15. Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M., and Mallick, B. K. Gene selection: a Bayesian variable selection approach. Bioinformatics, *19:* 90–97, 2003.

16. Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O-P, Borg, A., and Trent, J., Gene expression profiles in hereditary breast cancer. N. Engl. J. Med., *344:* 539–548, 2001.

17. Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med., *7:* 673–679, 2001.

18. Chien, J. T. On-line hierarchical transformation of hidden Markov models for speech recognition. IEEE Trans. Speech Audio Processing, *7:* 656–667, 1999.

Molecular Cancer Therapeutics

Binarization of Microarray Data on the Basis of a Mixture Model [1]

Xiaobo Zhou, Xiaodong Wang and Edward R. Dougherty

Mol Cancer Ther 2003;2:679-684.